NOTE TO USERS

This reproduction is the best copy available.

UMI®

University of Alberta

THEORETICAL AND PRACTICAL IMPROVEMENTS TO GEOSTATISTICAL ESTIMATION AND SIMULATION

by

Olena Babak ©

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Mining Engineering

Department of Civil and Environmental Engineering

Edmonton, Alberta
Fall 2008

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-46277-5
Our file Notre référence
ISBN: 978-0-494-46277-5

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

To my beloved husband Petro and son Stephan

Abstract

Geostatistics is a relatively new and rapidly growing discipline of engineering that provides a set of statistical and mathematical tools for generating numerical models of regionalized variables that help in data processing and decision making.

Over the years many geostatistical estimation and simulation techniques have been proposed. There are, however, a number of longstanding problems associated with these techniques. These problems include (1) correcting the "string effect" of kriging; (2) removing variance inflation of collocated simple cokriging; (3) improving practical implementation of collocated cokriging in the case of multiple secondary variables; (4) development of combined measure of local uncertainty, and (5) correcting multivariate simulation with correlated residuals to reproduce the target correlation between variables of interest.

These problems can result in poor estimates and poor distributions of uncertainty or systematic bias in the mean, variance and correlation between simulated realizations representing heterogeneity of the regionalized variables under study. The improved methods without these issues lead to numerical models with more realistic heterogeneity and improved uncertainty characterization.

Several new modeling techniques are proposed and developed within each of the five target areas: (1) distance constrained kriging and finite domain kriging; (2)

sequential Gaussian simulation with intrinsic collocated cokriging; (3) super-secondary approach; (4) overlap uncertainty technique; and (5) sequential Gaussian simulation that honors correlation at lag 0. Theoretical and/or practical developments and implementations are presented for each.

This work has resulted in significant progress over a wide range of topics. The techniques proposed in this thesis provide an important addition to geostatistical theory and practice. Future research and development directions are discussed and summarized in the conclusions.

Acknowledgements

It is a great pleasure to thank my co-supervisors, Dr. Clayton V. Deutsch and Dr. Oy Leuangthong, for all the advice, suggestions, discussions, help and support given during my studies at the University of Alberta. I am especially grateful to them for introducing me to the theory of geostatistics.

I would also like to thank all my colleagues at the Centre for Computational Geostatistics (CCG). I would especially like to thank Deepak Bhandari, David F. Machuca Mory and Mehran Hassanpour for their support and friendship.

The financial support for this research provided by the sponsors of the CCG, Alberta Ingenuity Fund, Andrew Stewart Memorial Graduate Prize, Petro-Canada Graduate Scholarship in Petroleum Engineering, Provost Doctoral Entrance Award and International Association for Mathematical Geology student grant is greatly appreciated.

I would like to acknowledge the encouragement of my parents, Ihor and Iryna Soroka, and sister Liliia Soroka. Finally and most importantly, I would like to thank to my husband Petro and son Stephan for their endless support, understanding and love.

Table of Contents

CHAPTER 1: Introduction……1
1.1. The Approach……2
1.3. Dissertation Outline……7

CHAPTER 2: Overview of Geostatistics……9
2.1. Random Variables……10
2.2. Stationarity and Ergodicity……10
    2.2.1. Stationarity……10
    2.2.2. Ergodicity……11
2.3. Measures of Spatial Variablity……12
2.4. Models of Coregionalization……15
    2.4.1. Linear Model of Coregionalization……15
    2.4.2. Markov Models……16
        2.4.2.1. Markov Model I……16
        2.4.2.2. Markov Model II……17
    2.4.3. Intrinsic Model of Coregionalization……18
2.5. Geostatistical Estimation……19
    2.5.1. Simple Kriging……20
    2.5.2. Ordinary Kriging……20
    2.5.3. Simple Cokriging……22
    2.5.4. Collocated Simple Cokriging……23
    2.5.5. Other Forms of Kriging……24
2.6. Geostatistical Simulation……25
    2.6.1. Matrix Simulation (LU Simulation)……26
    2.6.2. Sequential Gaussian Simulation……28
    2.6.3. Gaussian Cosimulation……29
    2.6.4. Review of Other Simulation Techniques……29
2.7. Discussion……30

CHAPTER 3: Correcting the String Effect……31
3.1. Introduction: String Effect……31
3.2. Methods for Fixing the String Effect……36
    3.2.1. Quick Fix……36
    3.2.2. Extend the String……36
    3.2.3. Use Simple Kriging……37
    3.2.4. Wrap the String……37
    3.2.5. Finite Domain Kriging of Deutsch (1994)……39

3.3. Distance Constrained Kriging.......41
    3.3.1. Methodology.......41
    3.3.2. Implementation.......42
        3.3.2.1. Program.......43
    3.3.3. Small Examples.......44
    3.3.4. Properties.......47
    3.3.5. Distance Constrained Kriging: Generalization to the Case of Multiple Strings.......48
3.4. Finite Domain Kriging.......49
    3.4.1. Methodology.......49
    3.4.2. Small Examples.......52
    3.4.3. Properties.......56
    3.4.4. Finite Domain Kriging: Generalization to the Case of Multiple Strings.....59
3.5. Comparison of Distance Constrained Kriging and Finite Domain Kriging.......60
3.6. Discussion.......63

CHAPTER 4: Uncertainty as the Overlap of Distributions.......65
4.1. Introduction.......65
4.2. Uncertainty as the Overlap of Distributions.......67
4.3. Example with Inverse Distance and Simple Kriging.......68
    4.3.1. Inverse Distance Interpolation.......69
    4.3.2. Uncertainty Overlap.......70
4.4. Case Study.......72
4.5. Discussion.......79

CHAPTER 5: Intrinsic Collocated Cokriging.......80
5.1. Introduction.......81
5.2. Simple Collocated Cokriging is not an Intrinsic Model.......82
5.3. Sources of Variance Inflation in Collocated Cokriging.......84
5.4. Intrinsic Collocated Cokriging: Examples.......87
    5.4.1. Example 1.......87
    5.4.2. Example 2.......94
5.5. Cokriging versus Collocated Cokriging.......98
    5.5.1. Difference in the Profiles of Weights.......99
    5.5.2. Difference in the Results of Estimation.......101
5.6. The Super Secondary Approach.......107
    5.6.1. Collocated Simple Cokriging with Multiple Secondary Data.......107
    5.6.2. Merging Multiple Secondary Variables.......108
    5.6.3. Proof of the Super Secondary Approach.......109
5.7. Extension of the Super Secondary Approach to Intrinsic Collocated Cokriging....115
    5.7.1. Intrinsic Collocated Cokriging with Multiple Secondary Data.......115
    5.7.2. Intrinsic Collocated Cokriging with Super Secondary Variable.......116
    5.7.3. Example.......116

5.8. Discussion.......119

CHAPTER 6: Multivariate SGS Honoring a Correlation Matrix.......120

6.1. Multivariate Sequential Gaussian Simulation.......120
6.2. Reproducing the Correlation Matrix between Variables at Lag 0.......123
6.3. Unconditional Multivariate SGS: Examples.......125
    6.3.1. Example 1.......125
    6.3.2. Example 2.......128
        6.3.2.1. Multivariate SGS for Z_1, Z_2 and Z_3.......130
        6.3.2.2. Multivariate SGS for Z_1, Z_2, Z_3 and Z_4.......133
        6.3.2.3. Multivariate SGS for All Five Variables.......135
6.4. Conditional Multivariate SGS: Example.......139
6.5. Local Correction.......143
6.6. Discussion.......146

CHAPTER 7: Discussion and Conclusions.......147

7.1. Summary.......148
    7.1.1. Correcting the String Effect.......148
    7.1.2. Uncertainty as the Overlap of Distributions.......150
    7.1.3. Intrinsic Collocated Cokriging.......151
    7.1.4. Multivariate SGS Honoring a Correlation Matrix.......152
7.2. Future Work.......154
    7.2.1. Improving LU/P-field Simulation.......154
    7.2.2. Testing for a Multivariate Gaussian Distribution.......155
    7.2.3. Variogram Upscaling.......155
    7.2.4. Accounting for the Uncertainty in Mean.......156

Bibliography.......158

Appendices

A Statistical Approach to Deterministic Inverse Distance Interpolation.......167

A.1. Introduction.......167
A.2. Background: Kriging versus Inverse Distance Interpolation.......169
A.3. Inverse Distance Interpolation.......171
A.4. Statistical Formalism.......172
    A.4.1. Stationarity.......172
    A.4.2. Mean and Variance of the Inverse Distance Estimator for the Stationary Domain.......172
A.5. Sensitivity of the Inverse Distance Weighted Interpolation to the Number of Data: Example.......174
    A.5.1. Data.......174
    A.5.2. Estimation.......174
    A.5.3. Sensitivity.......178

A.6. Inverse Distance with Locally Varying Parameters: Small Example......178
A.7. Comparison of the Inverse Distance with Locally Varying Parameters with Kriging and Inverse Distance Interpolation......182
A.8. Discussion......187

B Direct Upscaling of Variograms and Cross Variograms for Scale Consistent Geomodeling......189
B.1. Introduction......190
B.2. Regularization......192
B.3. Variogram Scaling Laws......192
B.4. Direct Variogram Upscaling......195
B.5. Linear Model of Coregionalization at a Block Support......197
B.6. Calculating Average Covariance $\overline{C}(V,V_h)$ and Average Variogram $\bar{\gamma}(V,V_h)$......200
B.7. Example: Scaling Laws vs. Direct Variogram Upscaling......201
B.8. Discussion......203

C Modeling Local Uncertainty accounting for the Uncertainty in Data......204
C.1. Simple Kriging......205
C.2. Calculating Mean and Variance of the Local Conditional Distributions accounting for the Uncertainty in Data......206
C.3. Small Examples......208
    C.3.1. Example 1......208
    C.3.2. Example 2......211
C.4. Discussion......213

List of Tables

3.1. The value in the left hand side of inequality (3.41) as a function of the number of data used in the finite domain simple kriging estimation....................................57

3.2. The value in the left hand side of inequality (3.41) as a function of the number of data used in the finite domain ordinary kriging estimation....................................59

3.3. Performance of OK vs. FDOK and DCOK in jackknife....................................61

A.1. Results of the cross validation for all 310 data in the study domain obtained based on the inverse distance interpolation with 3, 6, 12 and 24 data and power exponent $p = 1, 2, 3, 4$ and 6 and based on the local inverse distance interpolation with optimal parameters....................................180

A.2. Variogram model for the omnidirectional variograms of the nine exhaustive data sets....................................182

A.3. Mean square error obtained in estimation of the nine study areas with exhaustive data based on 100 data on a regular grid....................................186

A.4. Mean square error obtained in estimation of the nine study areas with exhaustive data based on 100 randomly selected data....................................186

C.1: Data locations and values....................................208

C.2. Effect of σ_i^2's on the local uncertainty distribution....................................209

C.3. Effect of μ_i's on the local uncertainty distribution....................................210

C.4. Theoretically-derived approach vs. Monte-Carlo simulation: Variance of the local uncertainty distribution....................................211

List of Figures

3.1. Finite strings of data as a result of stratigraphic boundary truncation (redrawn from Deutsch, 1993)...33

3.2. Finite strings of data as a result of a limited search (redrawn from Deutsch, 1993)..33

3.3. Profiles of ordinary kriging weights obtained for the estimation point located on a distance equal to the range of correlation from the string. Results are calculated based on a spherical variogram model with a nugget effect of 0% (a), 25% (b) and 75% (c) using string of 7 data...34

3.4. Profiles of the ordinary kriging weights obtained for the estimation point located on the distance equal to the range of correlation (a), 50% of the range of correlation (b) and 20% of the range of correlation (c). Results are calculated based on a spherical variogram model with a nugget effect of 20% using a string of 7 data.................................35

3.5. Profiles of the simple kriging weights obtained for the estimation point located on the distance equal to the range of correlation (a), 50% of the range of correlation (b) and 20% of the range of correlation (c). Results are calculated based on a spherical variogram model with a nugget effect of 20% using a string of 7 data................................38

3.6. Profiles of the distance constrained simple kriging (solid line) and simple kriging (dashed line) weights. Results shown for estimation locations: a) (1, 7); b) (1.8, 7); c) (2.8, 7); and d) (3.8, 7) are obtained using a spherical variogram model with a range of 20 based on a string of 7 data located at (1,0), (2,0), (3,0), (4,0), (5,0), (6,0) and (7,0), respectively. The closest data in a string is denoted by a dark circle........................45

3.7. Profiles of the distance constrained ordinary kriging (solid line) and ordinary kriging (dashed line) weights. Results shown for estimation locations: a) (1, 7); b) (1.8, 7); c) (2.8, 7); and d) (3.8, 7) are obtained using a spherical variogram model with a range of 20 based on a string of 7 data located at (1,0), (2,0), (3,0), (4,0), (5,0), (6,0) and (7,0), respectively. The closest data in a string is denoted by a dark circle........................46

3.8. Profiles of the finite domain simple kriging (solid line) and simple kriging (dashed line) weights. Results shown for estimation locations: a) (1, 7); b) (1.8, 7); c) (2.8, 7); and d) (3.8, 7) are obtained using a spherical variogram model with a range of 20 based on a string of 7 data located at (1,0), (2,0), (3,0), (4,0), (5,0), (6,0) and (7,0), respectively. The closest data in a string is denoted by a dark circle..53

3.9. Profiles of the finite domain ordinary kriging (solid line) and ordinary kriging (dashed line) weights. Results shown for estimation locations: a) (1, 7); b) (1.8, 7); c) (2.8, 7); and d) (3.8, 7) are obtained using a spherical variogram model with a range of 20 based on a string of 7 data located at (1,0), (2,0), (3,0), (4,0), (5,0), (6,0) and (7,0), respectively. The closest data in a string is denoted by a dark circle........................54

3.10. Change in the structure of the simple kriging weights (dashed lines) with respect to the number of closest data in a string used for estimation of location (3.8,7). Results are shown for: a) 1 data; b) 2 data; c) 4 data; and d) 7 data are obtained using spherical variogram model with the range of correlation 20. Finite domain simple kriging weights calculated based on all 7 data are shown in solid line. String of 7 data is located at (1,0), (2,0), (3,0), (4,0), (5,0), (6,0) and (7,0), respectively. The closest data in a string is denoted by a dark circle..55

3.11. Change in the structure of the finite domain simple kriging weights with respect to the number of closest data in a string used for estimation of location (100,7). Results are obtained using spherical variogram model with the range of correlation 500. The string of 3000 data is located at (1,0), (2,0), ..., (3000,0), respectively..............................58

3.12. Change in the structure of the finite domain ordinary kriging weights with respect to the number of closest data in a string used for estimation of location (100,7). Results are obtained using spherical variogram model with the range of correlation 500. The string of 3000 data is located at (1,0), (2,0), ..., (3000,0), respectively..............................58

3.13. Locations of 180 wells with data (a) together with bitumen distribution (b). Estimation wells are shown in dark circles..62

3.14. Experimental variogram and its theoretical fit in the three directions of major continuity for the normal score transformed bitumen from 90 estimation wells.........62

3.15. Maps of the differences between OK and DCOK (a) and OK and FDOK (b) for the slice at 265 m of the 3D model for the normal score transformed bitumen.................63

4.1. Schematic representation of the example results for the two local uncertainty at location $\mathbf{u}_0$. Solid and dashed lines represent local uncertainty obtained by two different estimation approaches; dark area is an overlap..68

4.2. Schematic representation of the two example results for the four local uncertainty at location $\mathbf{u}_0$. Solid dashed, doted and dash-dot lines represent local uncertainty obtained by four different estimation approaches; dark area is an overlap..........................69

4.3. Location map of 10 data (a) and their distribution (b). Points *A* and *B* represent two estimation locations of interest..71

4.4. (Scaled) overlap uncertainty estimator together with inverse distance and simple kriging local uncertainty models for estimation location *A* (a); and *B* (b). Dash-dot, dashed and solid lines represent local uncertainty obtained by simple kriging, inverse distance and overlap uncertainty approaches...71

4.5. Location map of 100 data from file 'cluster.dat' (a) and their distribution (b)......73

4.6. Accuracy plots for inverse distance (a), simple kriging (b), and overlap uncertainty estimator (c). Results are obtained in crossvalidation of 100 normal score transformed primary data from file 'cluster.dat'...74

4.7. Crossplots between: P_{10} (a) and P_{90} (b) of the simple kriging and overlap uncertainty estimator and crossplots between P_{10} (c) and P_{90} (d) of the inverse distance and overlap uncertainty estimator for 100 normal score transformed primary data from file 'cluster.dat'..75

4.8. (P_{10}, P_{90}) probability intervals obtained for the first 10 data in the 'cluster.dat' data set based on the simple kriging (dash-dot lines), inverse distance interpolation (dashed lines) and overlap uncertainty estimator (solid lines). Medians (P_{50}) for each of the three considered approaches are shown by dots...76

4.9. Crossplots between the variance of the local conditional distributions (smoothing effect) of the: inverse distance and simple kriging (top left); inverse distance and overlap uncertainty estimator (top right) and simple kriging and overlap uncertainty estimator (bottom) obtained for the 100 data of the file 'cluster.dat'..................................77

4.10. Cross validation results for the inverse distance interpolation (a); simple kriging (b); and overlap uncertainty estimator (c) obtained for 100 normal score transformed primary data from file 'cluster.dat'..78

5.1. Distribution of the means of the secondary standard normal random variable *Y* for the 100 sequential Gaussian realizations (a); and distribution of the variances of the secondary standard normal random variable *Y* for the 100 sequential Gaussian realizations (b)..88

5.2. Distribution of the means of the primary standard normal random variable *Z* for the 100 sequential Gaussian realizations based on simple collocated cokriging (a); and distribution of the variances of the primary standard normal random variable *Z* for the 100 sequential Gaussian realizations based on simple collocated cokriging (b).........88

5.3. Distribution of the means of the primary random variable *Z* for the 100 sequential Gaussian realizations based on intrinsic collocated cokriging (a); and distribution of the variances of the primary random variable Z for the 100 sequential Gaussian realizations based on intrinsic collocated cokriging (b)..90

5.4. Distribution of the means (first column) and variances (second column) of the primary random variable Z for the 100 sequential Gaussian realizations based on intrinsic collocated cokriging for the secondary data configuration shown in Schematic 1 with secondary data separation distance in X and Y directions: 1 grid cell (a); 2 grid cells (b) and 5 grid cells (c), respectively...91

5.5. Reproduction of the secondary variable Y semivariogram by sequential Gaussian simulation...93

5.6. Reproduction of the primary variable Z semivariogram by the sequential Gaussian simulation with simple collocated cokriging...93

5.7. Reproduction of the primary variable Z semivariogram by sequential Gaussian simulation with intrinsic collocated cokriging...94

5.8. Reproduction of the correlation between primary and secondary random variables for sequential Gaussian simulation with intrinsic collocated cokriging.........................95

5.9. Locations of the 20 primary data (a) and their distribution (b); the crossplot between primary data and collocated secondary data (c) and the map of exhaustive secondary data (d). The data are in Gaussian units..96

5.10. Example sequential Gaussian realization obtained based on simple collocated cokriging (a) and intrinsic collocated cokriging (b)..97

5.11. Distribution of the means of the primary random variable for the 100 sequential Gaussian realizations based on simple collocated cokriging (a); and distribution of the variances of the primary random variable for the 100 sequential Gaussian realizations based on simple collocated cokriging (b)..98

5.12. Distribution of the means of the primary random variable for the 100 sequential Gaussian realizations based on intrinsic collocated cokriging (a); and distribution of the variances of the primary random variable for the 100 sequential Gaussian realizations based on intrinsic collocated cokriging (b)...99

5.13. Reproduction of the correlation between primary and secondary random variables for sequential Gaussian simulation with simple collocated cokriging (a) and sequential Gaussian simulation with intrinsic collocated cokriging (b)............................100

5.14. Locations of 12 primary data (a) and their distribution (b), the crossplot between primary data and collocated secondary data (c) and the map of exhaustive secondary data (d)..102

5.15. Study domain with conditioning data (circles) and the estimation location (10, 10) (asterisk) (a); Primary data weights as a function of the ordered conditioning data, ordered according to the closeness to the estimation location (b) and secondary data weights as a function of the ordered conditioning data, zero stands for the estimation location (c)....103

5.16. Study domain with conditioning data (circles) and the estimation location (35, 35) (asterisk) (a); Primary data weights as a function of the ordered conditioning data, ordered according to the closeness to the estimation location (b) and secondary data weights as a function of the ordered conditioning data, zero stands for the estimation location (c)....104

5.17. The maps of estimates (a) and estimation variances (b) obtained based on collocated simple cokriging (top), simple cokriging with the intrinsic correlation model (middle) and simple cokriging with the linear model of corregionalization (bottom)....106

5.18. The maps of the difference in means (a) and variances (b) for collocated simple cokriging and simple cokriging with the intrinsic correlation model (top) and for simple cokriging with the linear model of corregionalization and simple cokriging with the intrinsic correlation model (bottom)....127

6.1. The histogram of the coefficients b_{12} (a) and a_{12} (b) obtained in 100 fully dependent and fully independent, respectively, multivariate SGS....127

6.2. Distribution of the correlation coefficients between Z_1 and Z_2 obtained by the corrected multivariate SGS (a); and distribution of the correlation coefficients between Z_1 and Z_2 obtained by conventional approach (b)....127

6.3. The variogram reproduction for Z_1 (a) and Z_2(b) obtained in the corrected multivariate SGS....128

6.4. Distributions of coefficients b_{12} (a), b_{13} (c), b_{23} (e), a_{12} (b), a_{13} (d), and a_{23} (f) obtained by 100 fully dependent and fully independent, respectively, multivariate SGS....131

6.5. Distribution of the correlation coefficients ρ_{12}, ρ_{13} and ρ_{23} obtained by the corrected multivariate SGS (a) and by conventional approach (b)....132

6.6. Variogram reproduction for Z_1, Z_2, and Z_3 obtained in the corrected multivariate SGS....134

6.7. Correlation matrix between Z_1, Z_2, Z_3, and Z_4 reproduced by the corrected multivariate SGS (a) and by conventional approach (b)....134

6.8. The mismatch in the reproduced correlation matrix between Z_1, Z_2, Z_3, and Z_4 obtained by the corrected multivariate SGS (a) and by conventional approach (b)......135

6.9. Variogram reproduction for Z_1, Z_2, Z_3 and Z_4 obtained in the corrected multivariate SGS..........136

6.10. Correlation matrix between Z_1, Z_2, Z_3, Z_4, and Z_5 reproduced by the corrected multivariate SGS (a) and by conventional approach (b)..........138

6.11. The mismatch in the reproduced correlation matrix between Z_1, Z_2, Z_3, and Z_4 obtained by the corrected multivariate SGS (a) and by conventional approach (b)......138

6.12. Locations of the 20 primary data (a) and their distribution (b); the crossplot between primary data and collocated secondary data (c) and the distribution of the secondary data (d). The data are in Gaussian units..........140

6.13. Distributions of coefficients b_{12} (a) and a_{12} (b) obtained in conditional multivariate SGS example..........141

6.14. Distribution of the correlation coefficients between primary and secondary random variables obtained by the corrected multivariate SGS..........141

6.15. Variogram reproduction in the direction of major and minor continuity for primary (a) and secondary (b) random variables obtained in the corrected multivariate SGS....142

6.16. Cross variogram reproduction in the direction of major and minor continuity obtained in the corrected multivariate SGS..........143

A.1. Location map of 310 samples (a) together with their distribution (b)..........175

A.2. Experimental omnidirectional variogram (points) together with its variogram fit for 310 samples in the study domain..........175

A.3. Results of the inverse distance interpolation for the mean (a) and variance (b) of the local conditional distributions obtained based on 3 data with exponent value *p* equal to: 1 (top); 3 (middle) and 9 (bottom)..........176

A.4. Results of the inverse distance interpolation for the mean (a) and variance (b) of the local conditional distributions obtained based on 24 data with exponent value *p* equal to: 1 (top); 3 (middle) and 9 (bottom)..........177

A.5. The estimation variance for the inverse distance interpolator with exponent value of 1 ($p = 1$) as a function of the number of data for the slice at $X = 100$..........179

A.6. The estimation variances for the inverse distance interpolator obtained based on 3 data (a) and the estimation variances for the inverse distance interpolator obtained based on 24 data (b) as a function of the power exponent for the slice at $X = 100$............179

A.7. Result of the optimal local inverse distance interpolation for the mean (a) and variance (b) of the local conditional distributions; optimal power exponent (c) and optimal number of data (d) for all estimation locations in the study domain...............181

A.8. Standardized omnidirectional variograms and their theoretical fits for the nine exhaustive data sets...183

A.9. Locations (asterisk) of 100 points in a) data set selected on a regular grid and in b) data set selected with random clustered pattern. Dots denote other locations with exhaustive data in the study domain of 50 by 50 meters....................................184

A.10. Example of ominidirectional variograms with their fits obtained based on 100 randomly selected data for data set 1 (a) and 100 data on a regular grid for data set 5 (b) and data set 6 (c)...185

B.1. Comparison of the upscaled variograms obtained using scaling laws (dashed lines) and direct variogram upscaling approach (solid lines) for the block support of 2m, 5m and 10m..202

B.2. Comparison of the upscaled variograms obtained using scaling laws (dashed lines) and direct variogram upscaling approach (solid lines) for the block support of 2m and 5m for lag distances up to 10m...202

C.1. Data configuration for Example 1..209

C.2. Variance of the local conditional distributions accounting for the uncertainty in the data obtained based on a single structured spherical variogram with nugget effect of 0 and range of continuity 5 (a) and 10 (b) : case 3 (top), case 4 (middle) and case 5 (bottom)...212

CHAPTER 1

Introduction

Geostatistics is a relatively new and rapidly growing branch of applied statistics that focuses on the geologic nature of the data, spatial relationship between observations and their volume support and precision (Deutsch, 2002). It consists of a set of statistical and mathematical tools for generating numerical models of regionalized variables that help in data processing and decision making.

Over the years, geostatistics has become a powerful tool in many areas of natural resources characterization. It is widely used to quantify uncertainty in energy and mineral resources (Journel and Huijbregts, 1978; Chiles and Delfiner, 1999). Other applications consist of generating input for flow simulation (Deutsch, 2002) and calculating the likelihood of exceeding critical threshold in contamination studies (Kyriakidis and Journel, 2001).

Geostatistical procedures rely on kriging-based techniques for optimal estimation and to model local conditional distributions. Kriging uses the spatial correlations provided by the variogram to calculate the weights that are applied to the sample values surrounding an unsampled location. The weights obtained from the kriging minimize the estimation variance and account for the spatial correlation between the surrounding samples and the estimation location (that is, closeness to the estimation location) and between samples themselves (that is, data redundancy).

Stochastic simulation is an essential element of modern geostatistics. It represents a powerful tool for the description of phenomena that cannot be described deterministically due to their inherent complexity (Almeida and Journel, 1994).

Geostatistical simulation is performed by drawing from the local conditional distributions constructed by kriging (Journel and Kyriakidis, 2004). Simulation allows the construction of multiple realizations of geological heterogeneity that reproduce the local conditioning data, data histogram and spatial variability described by the variogram. Sequential Gaussian simulation is a commonly used geostatistical simulation approach. This approach is one of the simplest and is based on an assumption of the multivariate Gaussian distribution. In the Gaussian framework, the kriging estimate and kriging estimation variance are exactly the mean and variance of the local conditional Gaussian distributions.

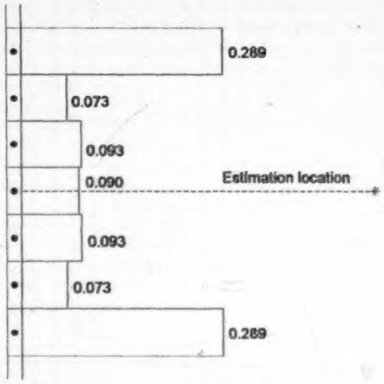
Geostatistical techniques, in general, focus on issues such as reproduction of the target statistics, uncertainty modeling and decision making after application of a transfer function. Kriging and simulation techniques are well known and applied in practice. However, some of these techniques suffer from problems such as variance inflation, biased reproduction of multivariate correlations, biased estimation or difficult implementation when there are multiple secondary data. There is a need for research to correct these problems.

1.1. The Approach

This research develops theoretically sound and practical methodologies and techniques for improved geostatistical modeling. The results provide important practical tools for geostatisticians and engineers. The proposed methodologies and techniques specifically address the following five important longstanding problems in geostatistics:

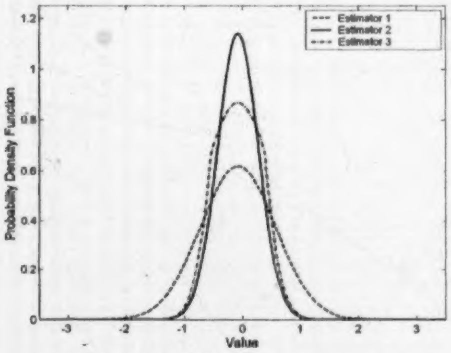
(1) *String Effect of Kriging*

There are two implicit assumptions behind kriging: stationarity and ergodicity. Stationarity is a decision to pool data for common analysis and a decision of the location-independence of the random function probability distribution and all its moments over the study domain. Stationarity is usually addressed by considering trends and locally varying parameters of estimation. Ergodicity is almost never addressed; it is fundamental to kriging. Ergodicity means that spatial averages (formed from averaging responses over many locations) are equivalent to probabilistic averages (formed from averaging over multiple replicates of the spatial process). A consequence of this assumption is that end samples in strings of data receive large weights in kriging, see figure in the left. This figure shows an example profile of ordinary kriging weights obtained for the estimation point located to the right of the string. Results are calculated based on a spherical variogram model with a nugget effect of 0% using a string of 7 data. Such weighting is theoretically valid; however, poor estimates and poor distributions of uncertainty arise when the end samples are unusually high or low. Many geological settings exhibit trends and such unusual grades at the contacts between geological domains. A number of ad-hoc corrections have been proposed, but none of them provide a constrained solution with a well defined measure of optimality.



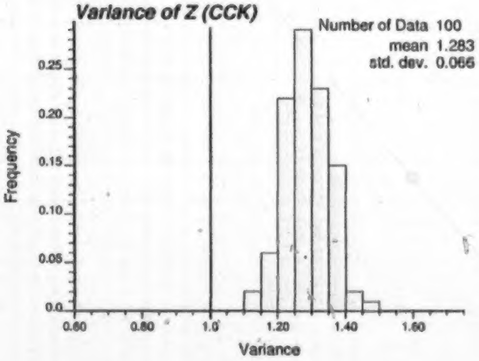
(2) *Obtaining a Combined ('Best') Measure of Local Uncertainty*

An important task in modern geostatistics is the assessment and quantification of local resource and reserve uncertainty. There are many different methods to build models of the local uncertainty including kriging, cokriging and inverse distance interpolation. Each of these techniques usually leads to different results, see figure on the left. This figure shows an example result obtained for the local uncertainty using three different estimation techniques including kriging and inverse distance. Due to difference in results, the modeler needs to choose the *best* approach. *Best* is theoretically optimal, has the greatest fidelity with the data, is the simplest to apply and so forth. In practice,



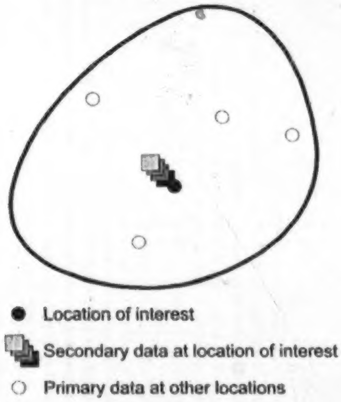
however, different techniques are *best* in different senses. In particular, kriging is a statistically optimal interpolator in the sense that it minimizes estimation variance when the variogram (measure of spatial continuity of the variable under study) is known and under the assumption of stationarity. In practice, these conditions are never satisfied. Inverse distance weighting, on the other hand, is a simple technique; it does not require prior information (variogram) to be applied to spatial prediction. There are many other techniques. A method for merging the uncertainties predicted by different models to obtain a combined measure of uncertainty that, ideally, possesses the good features of each alternative is of great practical interest.

(3) *Variance Inflation of Collocated Simple Cokriging*



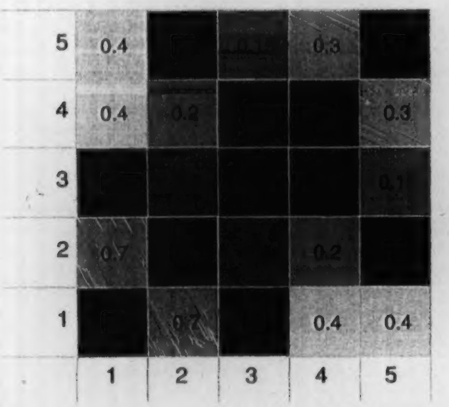
Sequential Gaussian simulation (SGS) with collocated cokriging is a widely used geostatistical simulation approach to integrate seismic data and to cosimulate multiple variables. This approach is based on a Markov-type hypothesis whereby collocated secondary data are assumed to screen further away data of the same type (Journel, 1999), that is, weighting the datum collocated with location being estimated is sufficient – nearby secondary data bring no new additional information. Sequential Gaussian simulation with collocated cokriging is popular because it is simple; the correlation coefficient between the primary variable being modeled and secondary data is the sole additional statistic required to integrate the secondary data. Collocated cokriging, however, has a longstanding problem with variance inflation that leads to a systematic bias in the mean and variance of the simulated realizations, see figure above left. This figure shows an example result for the distribution of variances of a primary standard normal random variable Z for 100 sequential Gaussian realizations based on simple collocated cokriging. Results are in Gaussian units. The target variance of 1 (shown by a dark vertical line) is clearly not met; the inflation of variance is more than 28%.

(4) *Accounting for Multiple Secondary Attributes in Collocated Cokriging*


● Location of interest
Secondary data at location of interest
○ Primary data at other locations

There often exist many secondary data that must be considered in geostatistical reservoir modeling including multiple seismic attributes, geological trends and structural controls. It is essential that all secondary data be accounted for when estimating a primary variable (e.g., porosity, water saturation, etc.) with the precision warranted by that secondary data type. The well data with primary variable are usually widely spaced and sparse, see schematic figure on the left. The secondary data are measured over the entire domain. There are a number of techniques to construct a distribution of uncertainty at an unsampled location using multiple well data at other locations and multiple secondary data. Cokriging is the standard technique in geostatistics to account for multiple data types. The most common variant of cokriging is collocated cokriging. Implementations of collocated cokriging are often limited to a single secondary variable. Practitioners often choose the most correlated or most relevant secondary variable. Improved models would be constructed if multiple variables were accounted for simultaneously.

(5) Correcting Multiple Univariate Sequential Gaussian Simulation with Correlated Residuals



When dealing with several attributes in spatial modeling it is desirable to reproduce the collocated correlation among them. An example target global collocated correlation matrix between four random variables is shown on the left. Reproduction of the correlation between multiple variables can be achieved by using variants of cokriging, in particular, simple cokriging. Simple cokriging, however, requires fitting a linear model of coregionalization – an intractable task in the case of many variables. There is, however, another more attractive and simpler alternative to cokriging. This alternative is multiple univariate sequential simulation with correlated residuals. This simulation approach is based on performing a matrix simulation with LU decomposition of the correlation matrix at each step of sequential simulation. Modeling of each random variable is performed independently. Thus, this type of simulation is primarily aimed at simulation of multiple primary variables. In simulation the data on each individual random variable is used to calculate the mean and variance of the local conditional distribution for that variable using simple kriging; no inference of the joint model for spatial continuity is needed. Multiple univariate SGS with correlated residuals has many desirable features including reproduction of the target mean and variance as well as target variogram models. However, an unfortunate feature of this technique is that while it targets reproduction of the correlation matrix between random variables of interest, acceptable reproduction of this correlation matrix is rarely observed in practice. The target correlation matrix is not reproduced because of conditioning to local data and a combination of the variable ordering and the sequential/LU decomposition.

Several new approaches were developed to deal with these five important problems. These approaches include: (1) distance constrained kriging and finite domain kriging for correcting the string effect; (2) sequential Gaussian simulation with intrinsic collocated cokriging as a solution to the variance inflation problem of collocated simple cokriging; (3) super secondary approach for merging multiple secondary variables into one variable for collocated cokriging; (4) overlap uncertainty approach for combining alternate distributions of local uncertainty; and (5) corrected multiple univariate sequential Gaussian simulation with correlated residuals.

All developed methods were coded. The developed code is publicly available. The format is similar to the Fortran code in GSLIB (Deutsch and Journel, 1998). Some of the programs were modified from GSLIB programs; which is acknowledged.

1.2. Dissertation Outline

Chapter 2 discusses the theoretical framework used in this dissertation with a short overview of the basic concepts in geostatistics. The concept of random variables is introduced. The assumptions of stationarity and ergodicity are explained. Measures of spatial variability are defined and models of coregionalization are provided. Classical geostatistical estimation and simulation methods are reviewed and discussed.

Chapter 3 focuses on the string effect of kriging. Section 3.1 defines and illustrates the string effect of kriging as well as presents the motivation for correcting it. Section 3.2 reviews the methods commonly applied for correcting the string effect. Sections 3.3 and 3.4 present two new methods for correcting the kriging weights to avoid large weights to data at the end of strings. These methods for resource estimation in a finite domain are distance constrained kriging and finite domain kriging. The methods are explained in detail and illustrated with small examples. Section 3.5 compares the two proposed kriging methods with a case study. A brief discussion of the results and the methods is presented in Section 3.6.

Chapter 4 focuses on combining local distributions of uncertainty obtained from different data sources or interpolation techniques. Section 4.1 reviews results of different comparative studies aimed at finding the best interpolation technique and explains the need for a combined method for local uncertainty calculation. Section 4.2 presents a new approach, referred to as the overlap uncertainty, for combining alternate conditional distributions of uncertainty. Application of the overlap uncertainty approach is shown in Section 4.3 by combining inverse distance interpolation results with simple kriging. Further analysis of the overlap uncertainty approach is conducted in Section 4.4 based on a small case study. A brief discussion of the results and the method is presented in Section 4.5.

Chapter 5 investigates the problem of variance inflation in collocated cokriging. Section 5.1 reviews alternative approaches for collocated cokriging. Section 5.2 investigates the theoretical justification for selecting only one auxiliary sample (collocated) for cokriging in the case of an intrinsic correlation model. Section 5.3 explains the reasons underlying variance inflation in collocated cokriging. Then intrinsic collocated cokriging to solve the variance inflation problem is presented. The proposed method is shown to remove systematic bias in histogram reproduction from conventional Markov models for collocated cokriging. Example applications are given in Section 5.4. Section 5.5 compares intrinsic collocated cokriging, collocated cokriging and cokriging with a linear model of coregionalization. Comparison is made in terms of the difference in the cokriging weight profiles and results of estimation. Section 5.6 presents an approach for improved implementation of collocated cokriging in the case of multiple secondary data. The theoretical validity of this approach is proven. Section 5.7 extends the proposed approach to intrinsic collocated cokriging. Finally, a brief discussion about the results and the methods is presented in Section 5.8.

Chapter 6 investigates multiple univariate sequential Gaussian simulation with correlated residuals. Section 6.1 reviews options for multivariate Sequential Gaussian Simulation. Section 6.2 presents a correction to multiple univariate Sequential Gaussian Simulation with correlated residuals to reproduce the correlation between random variables at lag 0. Sections 6.3 and 6.4 show application of the developed correction technique in several small examples and a case study. Section 6.5 presents an extension of the multiple univariate Sequential Gaussian Simulation to reproduce locally correlations between random variables. A brief discussion of the results and a correction method is presented in Section 6.6.

Additionally, this thesis includes several appendices with important related research. **Appendix A** proposes a statistical formalism for inverse distance interpolation. The proposed formalism is based on the assumption of stationarity and is aimed at providing the estimation variance at the unsampled locations as a measure of accuracy. A general approach to find the optimal exponent value and the optimal number of neighboring points to be used in estimation is also provided. **Appendix B** presents direct size scaling of variograms. Comparison of the direct approach with conventional variogram scaling laws is also given. Finally, **Appendix C** presents a general framework for integrating data uncertainty in spatial prediction.

CHAPTER 2

Overview of Geostatistics

Geostatistics is a branch of applied statistics that focuses on the geologic nature of the data, spatial relationship between observations and their volume support and precision (Deutsch, 2002). Geostatistics provides a set of statistical and mathematical tools for incorporating the spatial and temporal data coordinates in data processing (Goovaerts, 1997).

The development of geostatistics started in the 1960's and was driven by the problems in mining such as unbiased estimation of recoverable reserves. Problems in petroleum such as realistic heterogeneity models for unbiased flow predictions motivated the development of geostatistics from the mid 1980s through the late 1990s. Currently development of geostatistics is influenced by an increased need for more realistic geologic modeling and reliable uncertainty quantification (Deutsch, 2006).

Geostatistics is also applied to problems in forestry, fisheries, agriculture and environmental sciences (e.g., Jensen and Miller, 2005; Pople et al, 2007; Cockx et al, 2007). Geostatistical numerical models are used to spatially predict the variable of interest and to simulate its spatial features. Many geostatistical problems are concerned with reconciling data types at different scales (Deutsch, 2006).

This section presents a short overview of the basic concepts in geostatistics. More extensive discussion of these concepts can be found in the well known geostatistical books by David (1977); Journel and Huijbregts (1978); Isaaks and Srivastava (1989); Goovaerts (1997); Deutsch and Journel (1998); Chiles and Delfiner (1999); Wackernagel (2003) and Deutsch (2002) and many other journal and book publications.

2.1. Random Variables

A random variable Z is a variable that can take values according to a prescribed probability distribution. A spatially dependent random variable Z is denoted by $Z(\mathbf{u})$, where $\mathbf{u}$ is a location within the study domain A. An outcome of the random variable $Z(\mathbf{u})$ is denoted $z(\mathbf{u})$.

The uncertainty in the true value $z(\mathbf{u})$ at an unsampled location $\mathbf{u} \in A$ can be modeled using the cumulative probability distribution function of the random variable $Z(\mathbf{u})$,

$$F(\mathbf{u}; z) = Prob(Z(\mathbf{u}) \leq z). \tag{2.1}$$

This probability distribution function is a model of our lack of knowledge about the true value at the usampled location $\mathbf{u}$.

A set of several random variables $\{Z(\mathbf{u}_i), \quad \mathbf{u}_i \in A, \quad i = 1,\ldots,n\}$ is called a random function. Each of the random variables $Z(\mathbf{u}_i)$ has its own probability distribution; the random variables are dependent on each other. The dependence between random variables are characterized by the multivariate distribution also known as a spatial law (Goovaerts, 1997):

$$F(z(\mathbf{u}_1),\ldots,z(\mathbf{u}_n)) = Prob\,(Z(\mathbf{u}_1) \leq z(\mathbf{u}_1),\ldots,Z(\mathbf{u}_n) \leq z(\mathbf{u}_n)), \quad \mathbf{u}_i \in A, \quad i = 1,\ldots,n. \tag{2.2}$$

2.2. Stationarity and Ergodicity

2.2.1. Stationarity

Repetitive samples are needed to infer any statistic. Unfortunately, repetitive samples are not available in a spatial context. Most measurements cannot be repeated at the same location $\mathbf{u}$ to obtain a probability distribution of the random variable $Z(\mathbf{u})$. Stationarity is a decision to combine samples at other locations to obtain a model of the probability distribution. This amounts to assume the invariance of the random function and all its moments by translation over the domain A (Deutsch, 2002). In particular, first orderstationarity assumes that the mean of the variable of interest is constant throughout the domain A; second order stationarity assumes that the covariance between data depends only on the separation distance between data in the study domain A. That is,

$$\begin{aligned} E(Z(\mathbf{u})) &= m, \quad \forall \mathbf{u} \in A; \\ Cov(Z(\mathbf{u}), Z(\mathbf{u}+\mathbf{h})) &= C(\mathbf{h}), \quad \forall \mathbf{u} \in A. \end{aligned} \tag{2.3}$$

Stationarity is a property of the random function model; it is not a characteristic of the underlying spatial phenomenon. Since stationarity is not a hypothesis it cannot be tested; however, it can be judged inappropriate based on an understanding of the spatial phenomenon. Intrinsic stationarity is a less constraining assumption than second order stationarity. Intrinsic stationarity only assumes the existence of mean and variogram (the increments $Z(\mathbf{u})-Z(\mathbf{u}+\mathbf{h})$ are second order stationary). However, the covariance is not defined.

In practice, the decision of stationarity permits inference of the moments of the population given a set of samples $\{z(\mathbf{u}_i), \quad i=1,\ldots,n\}$. These moments are usually inferred using the experimental frequencies calculated from the data. Specifically, the stationary first order mean m is inferred from the global stationary univariate cumulative distribution function $F(z)$; the second order covariance $C(\mathbf{h})$ is inferred from the bivariate cdf of all z and z' sample data collected from all $\mathbf{u}$ to $\mathbf{u'}$ pairs of locations, respectively, approximately separated by the lag $\mathbf{h}$. Once these statistical parameters are quantified; they are used as input in geostatistical estimation and simulation algorithms to update the global stationary univariate cumulative distribution function (cdf) $F(z)$ to local conditional cumulative distribution functions (ccdf) $F(\mathbf{u};\ z|(n))$ using n sample data surrounding each unknown $\mathbf{u}$ location.

2.2.2. Ergodicity

A set of simulated values over the study domain is referred to as a realization of a random function model (Deutsch, 2002). The difference between the statistics of a simulated realization and the statistics of the random function model is called an ergodic fluctuation (Goovaerts, 1997). The ergodic theorem states that for a sufficiently large domain, the statistic of the realization should approximate the respective statistics of the random function model. In particular, a domain with size at least 8 times larger than the range of covariance function is considered to be large enough and, thus is called ergodic (Deutsch and Journel, 1998).

In practice, the ergodic fluctuations displayed by a simulated realization is controlled by the algorithm used to generate the realization; the density of conditioning data; the size of the simulated grid and the covariance parameters. Statistical parameters are inferred from sample information, not the complete population; therefore ergodic fluctuations account to some degree for the uncertainty about sample statistical parameters. Reduction or removal of ergodic fluctuations may lead to a false sense of certainty about the simulated features (Goovaerts, 1997).

2.3. Measures of Spatial Variablity

A variogram $2\gamma(\mathbf{h})$ is a measure of spatial variability and is an important statistic in geostatistical analysis. It measures the expected dissimilarity between two random variables $Z(\mathbf{u})$ and $Z(\mathbf{u}+\mathbf{h})$ separated by Euclidean distance **h**:

$$2\gamma(\mathbf{h}) = E\big(Z(\mathbf{u}) - Z(\mathbf{u}+\mathbf{h})\big)^2. \tag{2.4}$$

$\gamma(\mathbf{h})$ (half of the variogram) is referred to as the semivariogram (Journel and Huijbregts, 1978; Chiles and Delfiner, 1999).

The experimental variogram $2\hat{\gamma}(\mathbf{h})$ is calculated as the average of squared differences between data separated by a distance lag of **h**. In practice, angle and lag tolerances are defined so that reasonable number of data pairs $N(\mathbf{h})$ can be found approximately **h** apart (Deutsch and Journel, 1998):

$$2\hat{\gamma}(\mathbf{h}) = \frac{1}{N(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})}(z(\mathbf{u}_i) - z(\mathbf{u}_i+\mathbf{h}))^2 . \tag{2.5}$$

Reasonable choices in the number of lags, lag separation distance and tolerances usually helps to get a reliable estimate of the variogram, although this is not always feasible

(Cressie and Hawkins, 1980). Bad choices may result in noisy variograms that are not representative of the phenomenon under study.

In order to use experimental variograms in geostatistical estimation or simulation, they must be modeled (Gringarten and Deutsch, 2001). Variogram models must provide a valid measure of distance, thus, they must be positive semi-definite (Armstrong and Jabin, 1981). The positive semi-definiteness constraint ensures that the estimation variance will be non-negative. Several common variogram models that are know to fulfill this constraint are the Nugget Effect, Spherical, Gaussian and Exponential models. (Goovaerts, 1997).

The covariance is complementary to the variogram (Deutsch and Journel, 1998). The covariance function $C(\mathbf{h})$ is a measure of similarity between two random variables $Z(\mathbf{u})$ and $Z(\mathbf{u}+\mathbf{h})$ separated by Euclidean distance **h** (see also 2.3):

$$C(\mathbf{h}) = E\big(Z(\mathbf{u})Z(\mathbf{u}+\mathbf{h})\big) - m(\mathbf{u})m(\mathbf{u}+\mathbf{h}), \qquad (2.6)$$

where $m(\mathbf{u})$ and $m(\mathbf{u}+\mathbf{h})$ denote the means of $Z(\mathbf{u})$ and $Z(\mathbf{u}+\mathbf{h})$, respectively. If considering covariance at a lag $\mathbf{h} = 0$, then the definition of the covariance Equation (2.6) identifies the variance:

$$C(0) = Var(Z(\mathbf{u})).$$

Under the decision of stationarity, the semivariogram $\gamma(\mathbf{h})$, covariance function $C(\mathbf{h})$ and variance $C(0)$ are related as follows

$$\begin{aligned}
2\gamma(\mathbf{h}) &= E\big(Z(\mathbf{u}) - Z(\mathbf{u}+\mathbf{h})\big)^2 \\
&= E\big(Z(\mathbf{u})Z(\mathbf{u})\big) - 2E\big(Z(\mathbf{u})Z(\mathbf{u}+\mathbf{h})\big) + E\big(Z(\mathbf{u}+\mathbf{h})Z(\mathbf{u}+\mathbf{h})\big) \\
&= \big(E\big(Z(\mathbf{u})Z(\mathbf{u})\big) - m^2\big) - 2\big(E\big(Z(\mathbf{u})Z(\mathbf{u}+\mathbf{h})\big) - m^2\big) + \big(E\big(Z(\mathbf{u}+\mathbf{h})Z(\mathbf{u}+\mathbf{h})\big) - m^2\big) \\
&= 2C(0) - 2C(\mathbf{h}).
\end{aligned}$$

Or,

$$\gamma(\mathbf{h}) = C(0) - C(\mathbf{h}). \qquad (2.7)$$

When two or more variables are available, cross-variograms and cross-covariances are used to measure their spatial relationship. They measure how dissimilar and similar, respectively, are variables i and j ($i \neq j$) on average at two locations separated by Euclidean distance **h**,

$$\begin{aligned}2\gamma_{ij}(\mathbf{h}) &= E\big((Z_i(\mathbf{u})-Z_i(\mathbf{u}+\mathbf{h}))(Z_j(\mathbf{u})-Z_j(\mathbf{u}+\mathbf{h}))\big),\\ C_{ij}(\mathbf{h}) &= E\big(Z_i(\mathbf{u})Z_j(\mathbf{u}+\mathbf{h})\big)-m_i(\mathbf{u})m_j(\mathbf{u}+\mathbf{h}),\end{aligned} \tag{2.8}$$

where $m_i(\mathbf{u})$ and $m_j(\mathbf{u}+\mathbf{h})$ denote the means of $Z_i(\mathbf{u})$ and $Z_j(\mathbf{u}+\mathbf{h})$, respectively.

Under decision of stationarity, the cross-semivariograms $\gamma_{ij}(\mathbf{h})$, cross-covariance functions $C_{ij}(\mathbf{h})$, and cross-covariance between collocated variables i and j, $C_{ij}(0)$ are related as follows (Goovaerts, 1997):

$$\begin{aligned}2\gamma_{ij}(\mathbf{h}) &= E\big((Z_i(\mathbf{u})-Z_i(\mathbf{u}+\mathbf{h}))(Z_j(\mathbf{u})-Z_j(\mathbf{u}+\mathbf{h}))\big)\\ &= E\big(Z_i(\mathbf{u})Z_j(\mathbf{u})\big)-E\big(Z_i(\mathbf{u})Z_j(\mathbf{u}+\mathbf{h})\big)-E\big(Z_i(\mathbf{u}+\mathbf{h})Z_j(\mathbf{u})\big)+E\big(Z_i(\mathbf{u}+\mathbf{h})Z_j(\mathbf{u}+\mathbf{h})\big)\\ &= \big(E\big(Z_i(\mathbf{u})Z_j(\mathbf{u})\big)-m_im_j\big)-\big(E\big(Z_i(\mathbf{u})Z_j(\mathbf{u}+\mathbf{h})\big)-m_im_j\big)\\ &\qquad -\big(E\big(Z_i(\mathbf{u}+\mathbf{h})Z_j(\mathbf{u})\big)-m_im_j\big)+\big(E\big(Z_i(\mathbf{u}+\mathbf{h})Z_j(\mathbf{u}+\mathbf{h})\big)-m_im_j\big)\\ &= 2C_{ij}(0)-2C_{ij}(\mathbf{h}).\end{aligned}$$

Or,

$$\gamma_{ij}(\mathbf{h}) = C_{ij}(0)-C_{ij}(\mathbf{h}). \tag{2.9}$$

Note that Equation (2.9) is valid only if the cross-covariance is assumed symmetric in **h**, that is, if $C_{ij}(\mathbf{h}) = C_{ji}(\mathbf{h})$. This assumption, however, may not necessarily be true. There are a number of different possible causes for non-symmetric covariances ($C_{ij}(\mathbf{h}) \neq C_{ji}(\mathbf{h})$), including a well known lag effect (Journel and Huijbregts, 1978). A very simple example of a non-symmetric covariance is the cross-covariance calculated between a continuous and derivable variable $Z(\mathbf{u})$ and its derivative $Y(\mathbf{u})$. In this case, covariance is anti-symmetric, that is, $C_{ZY}(\mathbf{h}) = -C_{YZ}(\mathbf{h})$ (Chiles and Delfiner, 1999). Because the cross-covariance may be asymmetric, it is important to check if the assumption of symmetric covariance is realistic for any real application (Wackernagel, 2003).

In order to incorporate experimental cross-variograms in estimation or simulation, they must be modeled in a mathematically consistent manner with each other, and provide a measure of spatial correlation that respects physical laws and ensures positive estimation variances (Goovaerts, 1997).

2.4. Models of Coregionalization

To characterize the spatial relationship between M random variables, $Z_j(\mathbf{u}), \quad j = 1,\ldots,M, \quad \mathbf{u} \in A$, a matrix of covariance functions C containing covariances between sample points is needed. This matrix of covariance functions must be positive semi-definite. This will ensure that all variances are non-negative as well as the existence and uniqueness of solutions in kriging-based estimation and simulation.

There are several models for the matrix of covariance functions. These include the linear model of coregionalization, Markov models and intrinsic model of coregionalization. A short description of these models follows.

2.4.1. Linear Model of Coregionalization (LMC)

The LMC is the most common model for describing joint spatial continuity of multiple random variables. It is based on the assumption that each random variable $Z_j(\mathbf{u}), j = 1,\ldots,M$, can be expressed as a linear combination of the same K independent stationary random functions, $Y_k(\mathbf{u}), k = 1,\ldots,K$, each with zero mean and distinct covariance function $C_k(\mathbf{h})$:

$$Z_j(\mathbf{u}) = \sum_{k=1}^{K} a_{ik} Y_k(\mathbf{u}) + \mu_j,$$

where μ_j denotes the mean of the random variable Z_j (Journel and Huijbregts, 1978; Isaaks and Srivastava, 1998).

The LMC is given by the following system (Deutsch, 2002):

$$\gamma_{ij}(\mathbf{h}) = \sum_{k=0}^{K} b_{ij}^{k} \Gamma^{k}(\mathbf{h}), \quad i, j = 1,\ldots,M, \tag{2.10}$$

where the Γ^0 denotes the nugget effect; $\Gamma^i, i = 1,\ldots,K$, are distinct nested structures that make up the common pool of variogram models (spherical, exponential, etc.). All direct and cross variogram use the same variogram nested structures. The sill contribution parameters are allowed to change such that the $K+1$ $M \times M$ matrices of coefficients b_{ij}^k, $k = 0,1,\ldots,K$, are positive semi-definite.

The LMC (2.10) can be applied to any number of random variables. Several procedures for automatic fitting of the linear model of coregionalization have been developed (e.g., Goulard and Voltz, 1992; Lark and Papritz, 2003; Pelletier, et al, 2004), however, the LMC is rarely applied to more than four random variables.

2.4.2. Markov Models

Modeling an LMC is a complex task. Two Markov models simplify this task: Markov Model I and Markov Model II.

2.4.2.1. Markov Model I

The Markov Model I (MMI) assumes that the primary Z data prevails over collocated secondary Y data. Formally, it can be written as (Xu et al., 1992; Almeida and Journel, 1994; Goovaerts, 1997):

$$E(Y(\mathbf{u}) \mid Z(\mathbf{u}) = z, Z(\mathbf{u}+\mathbf{h}) = z') = E(Y(\mathbf{u}) \mid Z(\mathbf{u}) = z), \quad \forall \mathbf{h}, \forall z'. \qquad (2.12)$$

That is, dependence of the secondary variable on the primary is limited to the collocated primary datum. The cross covariance $C_{YZ}(\mathbf{h})$ under the Markov model I is given by:

$$C_{YZ}(\mathbf{h}) = b \cdot C_Z(\mathbf{h}), \quad \forall \mathbf{h} \qquad (2.13)$$

where $C_Z(\mathbf{h})$ is the covariance of Z; $b = \left(\sqrt{C_Y(0)} \middle/ \sqrt{C_Z(0)} \right) \rho_{YZ}(0)$, $C_Z(0)$ and $C_Y(0)$ are the variances of Z and Y; and $\rho_{YZ}(0)$ is the correlation coefficient of collocated Z and Y data.

If only collocated secondary data are used in estimation, then the primary variable variogram (or covariance) and the correlation coefficient between primary and secondary data are the only information required under the Markov Model I to obtain the joint model for the matrix of covariance functions (Almeida and Journel, 1994).

Note that the MM1 is a reasonable model of coregionalization if Z is defined on the same or a larger volume support than Y. The primary data Z screens the influence of further away data z'. In many practical applications, however, the primary data Z is defined on a smaller support than the secondary data $Y(\mathbf{u})$. In this case, the experimental cross-covariance $\hat{C}_{YZ}(\mathbf{h})$ tends to share the shape of the smooth secondary covariance model $C_Y(\mathbf{h})$ and, thus, Markov model I is inappropriate (Journel, 1999).

2.4.2.2. Markov Model II

The Markov Model II assumes that the secondary Y data prevails over the primary Z data. Formally, it can be written as (Journel, 1999; Shmaryan and Journel, 1999):

$$E\big(Z(\mathbf{u}) \mid Y(\mathbf{u}) = y, Y(\mathbf{u}+\mathbf{h}) = y'\big) = E\big(Z(\mathbf{u}) \mid Y(\mathbf{u}) = y\big), \quad \forall \mathbf{h}, \forall y' \tag{2.14}$$

That is, dependence of the primary variable on the secondary is limited to the collocated secondary datum. The cross covariance $C_{YZ}(\mathbf{h})$ under the Markov model II is given by:

$$C_{YZ}(\mathbf{h}) = b \cdot C_Y(\mathbf{h}), \quad \forall \mathbf{h} \tag{2.15}$$

where $C_Y(\mathbf{h})$ is the covariance of Y; $b = \left(\sqrt{C_Z(0)} \Big/ \sqrt{C_Y(0)} \right) \rho_{YZ}(0)$, $C_Z(0)$ and $C_Y(0)$ are the variances of Z and Y; and $\rho_{YZ}(0)$ is the correlation coefficient of collocated Z and Y data.

Although the Markov Model II results in a simple model for the cross-covariance (see (2.15)); the primary variable variogram must be modeled in addition to the secondary variable variogram. This variogram can be obtained as follows

$$C_Z(\mathbf{h}) = C_{ZY}(0) C_Y(\mathbf{h}) + (1 - C_{ZY}(0)) C_R(\mathbf{h}), \tag{2.16}$$

where $C_{ZY}(0)$ is the cross-covariance between collocated Z and Y data; $C_Z(\mathbf{h})$ is the covariance of Z; and $C_R(\mathbf{h})$ is any permissible covariance function (Journel, 1999).

Although modeling of MM2 is more demanding than MMI, its underlying hypothesis is more appropriate for a secondary variable defined on a large support volume, e.g., remote sensing data such as seismic.

2.4.3. Intrinsic Correlation Model (ICM)

The intrinsic correlation model is another multivariate covariance model that can be adopted for a covariance function matrix. It describes the relationships between the variables with the variance-covariance matrix B ($B(i,j)=b_{ij},\quad i,j=1,\dots,M$) and the relations between points in space by a spatial correlation function $\rho(\mathbf{h})$ as follows:

$$C^{\mathrm{ICM}}(\mathbf{h}) = B\cdot\rho(\mathbf{h}). \tag{2.17}$$

Note that the spatial correlation function $\rho(\mathbf{h})$ is the same for all variables. The model (2.17) is called the intrinsic correlation model because the correlation ρ_{ij} between any two collocated variables *i* and *j* is independent of the spatial scale (Wackernagel, 2003):

$$\frac{b_{ij}\rho(\mathbf{h})}{\sqrt{b_{ii}\rho(\mathbf{h})b_{jj}\rho(\mathbf{h})}} = \frac{b_{ij}}{\sqrt{b_{ii}b_{jj}}}\rho_{ij}. \tag{2.18}$$

In practice, the intrinsic correlation model specifies that the direct and cross covariance functions are all proportional to the same underlying spatial correlation function:

$$C_{ij}^{\mathrm{ICM}}(\mathbf{h}) = b_{ij}\rho(\mathbf{h}), \tag{2.19}$$

where the coefficients b_{ij} represent the variances ($i=j$) and covariances ($i \neq j$) between collocated variables. An intrinsic model can be defined in terms of variograms. Specifically, the intrinsic correlation model is a product of a positive coregionalization matrix B of coefficients b_{ij} and a variogram $\gamma(\mathbf{h})$, that is:

$$\Gamma^{\mathrm{ICM}}(\mathbf{h}) = B\gamma(\mathbf{h}). \tag{2.20}$$

When comparing a linear model of correlation (2.10) with an intrinsic correlation model (2.19) we note that the ICM can be viewed as a special case of the LMC. When all direct and cross covariance functions are proportional to the same underlying spatial correlation function, the linear model of correlation reduces to the intrinsic correlation model.

2.5. Geostatistical Estimation

One of the most important problems in the geosciences is the problem of spatial prediction. Spatial predictions are often required for planning, risk assessment, and decision-making. Typical applications include determining the profitability of mining an orebody, producing a reservoir, management of soil resources, soil properties mapping, pest management, designing a network of environmental monitoring stations, etc. (Weisz et al, 1995; Gotway et al., 1996; Moyeed and Papritz, 2002).

Spatial prediction techniques differ from classical statistical modeling in that they incorporate information on the geographic position of the sample data points (Journel and Huijbregts, 1978; Cressie, 1993). Spatial predictions describe a variety of responses over different spatial scales (Schloeder, et al., 2001). They provide a unique and smooth distribution of estimates that reproduce the sample points (conditioning data); spatial prediction techniques aim at local accuracy (Isaaks and Srivastava, 1989; Journel et al., 2000). The most common spatial prediction techniques calculate the estimates for a property by averaging nearby data. Weighting for each averaged data value is assigned either according to deterministic or statistical (spatial covariance) criteria. When a deterministic criterion is used, the measures of optimality are arbitrarily chosen (Borga and Vizzaccaro, 1997). When a statistical criterion is used, the field is considered as a random process and the optimality of the averaging method is determined in terms of minimizing the estimation variance. The methods obtained based on the statistical criterion are often referred to as kriging or kriging-based techniques.

Kriging is a well-proven methodology that provides the best linear unbiased estimate and its variance at the unsampled location. Because it returns the original data values, it is considered an exact technique (Deutsch, 2002). In theory, kriging is a statistically optimal interpolator in the sense that it minimizes estimation variance when the covariance or variogram is known and under the assumption of stationarity. The assumption of stationarity is relaxed in some types of kriging.

A short description of the most renowned flavors of geostatistical kriging-based techniques is presented next. These include simple and ordinary kriging, simple cokriging and collocated cokriging (see for reference Journel and Huijbregts, 1978; Cressie, 1993; Deutsch, 2002).

2.5.1. Simple Kriging

The simple kriging estimator predicts the value of the variable of interest $Z(\mathbf{u})$ at the estimation location $\mathbf{u}$ as a linear combination of neighboring observations $Z(\mathbf{u}_i), i = 1,\ldots,n(\mathbf{u})$, (Journel and Huijbregts, 1978):

$$Z^*_{SK}(\mathbf{u}) = \sum_{i=1}^{n(\mathbf{u})} \lambda_i(\mathbf{u}) Z(\mathbf{u}_i) + \left[1 - \sum_{i=1}^{n(\mathbf{u})} \lambda_i(\mathbf{u})\right] m, \tag{2.21}$$

where m denotes the stationary mean, $\lambda = (\lambda_1(\mathbf{u}),\ldots,\lambda_{n(\mathbf{u})}(\mathbf{u}))^T$ denotes the vector of the simple kriging weights calculated from the normal system of equations

$$\sum_{i=1}^{n(\mathbf{u})} \lambda_i(\mathbf{u}) Cov(Z(\mathbf{u}_i), Z(\mathbf{u}_j)) = Cov(Z(\mathbf{u}), Z(\mathbf{u}_j)), \quad j = 1,\cdots,n(\mathbf{u}), \tag{2.22}$$

where $Cov(Z(\mathbf{u}_i), Z(\mathbf{u}_j)), \quad i, j = 1,\cdots,n(\mathbf{u})$, denotes the data-to-data covariance values and $Cov(Z(\mathbf{u}), Z(\mathbf{u}_j)), \quad j = 1,\cdots,n(\mathbf{u})$, is the data-to-estimation point covariance values. The covariance function is calculated under stationarity through the semivariogram model $\gamma(\mathbf{h})$, see Equation (2.7).

Simple kriging is the best linear unbiased estimator, that is, it provides estimates with minimum error variance $\sigma^2_{SK}(\mathbf{u})$ in the least square sense given by

$$\sigma^2_{SK}(\mathbf{u}) = C(0) - \sum_{i=1}^{n(\mathbf{u})} \lambda_i(\mathbf{u}) Cov(Z(\mathbf{u}), Z(\mathbf{u}_j)), \tag{2.23}$$

where $C(0)$ is the stationary variance.

2.5.2. Ordinary Kriging

Ordinary kriging is a common variant of kriging. It is believed to be a robust and reliable method of data interpolation (Yamamoto, 2005). The ordinary kriging estimator differs from simple kriging in that it constraints the sum of all weights to be 1. Specifically, it provides a model for the value of the variable of interest at the estimation location **u** as the following linear combination of the neighboring observations $Z(\mathbf{u}_i), i = 1,\ldots,n(\mathbf{u})$, (Journel and Huijbregts, 1978)

$$Z^*_{OK}(\mathbf{u}) = \sum_{i=1}^{n(\mathbf{u})} \eta_i(\mathbf{u}) Z(\mathbf{u}_i), \tag{2.24}$$

where $\eta = (\eta_1(\mathbf{u}),\ldots,\eta_{n(u)}(\mathbf{u}))^T$ denote the vector of the ordinary kriging weights calculated from the following system of equations for the estimation location **u**,

$$\begin{gathered}\sum_{i=1}^{n(\mathbf{u})} \eta_i(\mathbf{u}) Cov(Z(\mathbf{u}_i), Z(\mathbf{u}_j)) + \mu_{OK}(\mathbf{u}) = Cov(Z(\mathbf{u}), Z(\mathbf{u}_j)), \quad j = 1,\cdots,n(\mathbf{u}), \\ \sum_{i=1}^{n(\mathbf{u})} \eta_i(\mathbf{u}) = 1,\end{gathered} \tag{2.25}$$

where $\mu_{OK}(\mathbf{u})$ is the Lagrange parameter and, as before, $Cov(Z(\mathbf{u}_i), Z(\mathbf{u}_j))$, $i, j = 1,\cdots,n(\mathbf{u})$, denotes the data-to-data covariance values and $Cov(Z(\mathbf{u}), Z(\mathbf{u}_j))$, $j = 1,\cdots,n(\mathbf{u})$, is the data-to-estimation point covariance values.

The ordinary kriging estimator is also an exact interpolator. However, ordinary kriging provides estimates with larger error variance $\sigma^2_{OK}(\mathbf{u})$ than simple kriging,

$$\sigma^2_{OK}(\mathbf{u}) = C(0) - \sum_{i=1}^{n(\mathbf{u})} \eta_i(\mathbf{u}) Cov(Z(\mathbf{u}), Z(\mathbf{u}_j)) - \mu_{OK}(\mathbf{u}) \geq \sigma^2_{SK}(\mathbf{u}), \tag{2.26}$$

where $C(0)$ is the stationary variance.

Note that ordinary kriging is usually preferred to simple kriging in practical applications. This is because ordinary kriging does not require knowledge or assume stationarity of the mean over the entire region of interest. Ordinary kriging accounts for unknown or locally varying mean by limiting the region of stationarity to within the neighbourhood centered at the location of interest. In general, it can be shown that OK estimates are larger than SK estimates in high-valued areas where the local mean is larger than the global mean; and OK estimates are smaller than SK estimates in lower-valued areas where the local mean is smaller than the global mean (Goovaerts, 1997).

2.5.3. Simple Cokriging

Simple cokriging (CSK) is a natural extension of simple kriging to the case when multivariate data is available (Vauclin et al, 1983; Wackernagel 1994; Goovaerts, 1997; Wackernagel, 2003). Simple cokriging allows estimating with data of the same type and auxiliary variables in the neighborhood. Specifically, the estimator is the following weighted linear combination of the mean of the variable of interest (m) and the data from different variables located at sample points in the neighborhood of the estimation location $\mathbf{u}$:

$$Z^*_{CSK}(\mathbf{u}) = m + \sum_{i=1}^{N} \sum_{\alpha=1}^{n_i(\mathbf{u})} \lambda_\alpha^i (Z_i(\mathbf{u}_\alpha) - m_i), \tag{2.27}$$

where the CSK weights $[\lambda_1^T, \ldots, \lambda_N^T]^T$ are found from a simple cokriging system given by

$$\begin{pmatrix} C^{11} & \cdots & C^{1j} & \cdots & C^{1N} \\ \vdots & \ddots & & & \vdots \\ C^{i1} & \cdots & C^{ij} & & C^{iN} \\ \vdots & & & \ddots & \vdots \\ C^{N1} & \cdots & C^{Nj} & \cdots & C^{NN} \end{pmatrix} \begin{pmatrix} \lambda^1 \\ \vdots \\ \lambda^i \\ \vdots \\ \lambda^N \end{pmatrix} = \begin{pmatrix} c^{1*} \\ \vdots \\ c^{i*} \\ \vdots \\ c^{N*} \end{pmatrix}, \tag{2.28}$$

where the left hand side covariance matrix is built up with square symmetric n_i by n_i blocks C^{ii} given by

$$C_{kl}^{ii} = Cov(Z_i(\mathbf{u}_k), Z_i(\mathbf{u}_l)), \quad k,l = 1,\ldots,n_i, \quad i = 1,\ldots,N,$$

on the diagonal and with rectangular n_i by n_j blocks $C_{ij} = C_{ji}^T$ given by

$$C_{kl}^{ij} = Cov(Z_i(\mathbf{u}_k), Z_j(\mathbf{u}_l)), \quad k = 1,\ldots,n_i, \quad l = 1,\ldots,n_j, \quad i,j = 1,\ldots,N,$$

off the diagonal. Note that the blocks C_{ij} contain either direct ($i = j$) or cross ($i \neq j$) covariances between sample points. The vectors c^{i*} contain the covariances with the variable of interest, for a specific variable of the set, between sample points and the estimation location. The vectors λ^i represent the weight attached to the sample of the i-th variable.

A joint model for the matrix of covariance functions (for example, a linear model of coregionalization) is required in order to perform simple cokriging. Thus, when M

different variables are considered, the covariance matrix in the left hand side of simple cokriging Equation (2.28) requires $\frac{M(M+1)}{2}$ or M covariance functions to be inferred depending on whether this matrix is symmetric (no lag effect) or not (lag effect) (see for reference Journel and Huijbrogts (1978) or Goovaerts (1997)). Such inference is demanding in terms of data and subsequent joint modeling; therefore, a more simple estimation technique called collocated simple cokriging is commonly employed.

2.5.4. Collocated Simple Cokriging

Collocated simple cokriging is a strategy where the neighborhood of the auxiliary variable is reduced to the estimation location only. The value of the auxiliary variable $Y(\mathbf{u})$ is said to be collocated with the variable of interest $Z(\mathbf{u})$ at the estimation location $\mathbf{u}$. The collocated simple cokriging estimator (in the case of only one secondary data type) is given by (Goovaerts, 1997):

$$Z^*_{CCSK}(\mathbf{u}) = m_Z + \lambda_0(Y(\mathbf{u}) - m_Y) + \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha (Z(\mathbf{u}_\alpha) - m_Z), \tag{2.29}$$

where collocated simple cokriging weights $[\lambda_Z^T, \lambda_Y]^T$ are found from the following system of equations

$$\begin{pmatrix} C_{ZZ} & c_{YZ} \\ c_{YZ}^T & C_Y(0) \end{pmatrix} \begin{pmatrix} \lambda_Z \\ \lambda_Y \end{pmatrix} = \begin{pmatrix} c_{ZZ} \\ C_{YZ}(0) \end{pmatrix}, \tag{2.30}$$

where C_{ZZ} is the left hand matrix of the simple kriging system of $Z(\mathbf{u})$ and c_{ZZ} is the corresponding right hand side covariance vector. The vector c_{YZ} contains the cross covariances between the $n(\mathbf{u})$ sample points of Z and the estimation location $\mathbf{u}$ with its collocated value $Y(\mathbf{u})$. $C_Y(0)$ is the variances of Y; and $C_{YZ}(0)$ is the cross-covariance between collocated Z and Y data.

The cross covariance c_{YZ} in system (2.30) is usually calculated using the Markov correlation Model I (Journel, 1999). Using the Markov model I for the cross covariance c_{YZ}, we can rewrite system (2.29) for the collocated simple cokriging weights $[\lambda_Z^T, \lambda_Y]^T$ as:

$$\begin{pmatrix} C_{ZZ} & bc_Z \\ bc_Z^T & C_Y(0) \end{pmatrix} \begin{pmatrix} \lambda_Z \\ \lambda_Y \end{pmatrix} = \begin{pmatrix} c_{ZZ} \\ C_{YZ}(0) \end{pmatrix}, \tag{2.31}$$

where $b = \left(\sqrt{C_Z(0)} \middle/ \sqrt{C_Y(0)} \right) \rho_{YZ}(0)$, $C_Z(0)$ is the variances of Z, and $\rho_{YZ}(0)$ is the correlation coefficient of collocated Z and Y data.

In order to perform collocated cokriging with Markov model I, the primary variable covariance function, the variance of the secondary data and the correlation coefficient between primary and secondary data are required. Retaining only the collocated secondary data, in general, does not affect the resulting estimate, since nearby data are usually very similar in values. However, it may affect the cokriging estimation variance. Cokriging variances are overestimated, oftentimes significantly. This causes serious problems in sequential simulation (Deutsch, 2002).

2.5.5. Other Forms of Kriging

There are many other kriging-based techniques. They either make different assumptions on the mean or require the data to be transformed prior to applying kriging. In particular, the following forms of kriging are also commonly applied:

- **Simple kriging with locally varying mean.** The global mean m in the simple kriging estimator is replaced by a local mean $m(\mathbf{u})$. (Goovaerts, 1997; Deutsch and Journel, 1998; Deutsch, 2002)
- **Universal kriging or kriging with a trend.** The mean is assumed to follow a particular regular (smooth) function, for example polynomial (Deutsch and Journel, 1998), logarithm (Brochu and Marcotte, 2003) or trigonometric functions (Seguret, 1989). The mean is fitted as a scaling of the trend function. Simple kiging is performed with residuals from the implicitly calculated mean values. (Huijbregts and Matheron, 1971; Olea, 1973; Armstrong, 1984; Deutsch and Journel, 1998)
- **Kriging with external drift.** The mean values are estimated as a linear function of a secondary variable such as seismic. The mean is fitted as a scaling of the

external drift variable. Simple kiging is performed with residuals from the implicitly calculated mean values. (Moinard, 1987; Galli and Meunier, 1987; Castelier, 1993; Goovaerts, 1997; Bourennane, et al, 2000; Rivoirard, 2002)

- **MultiGaussian kriging.** The data are transformed into the standard normal distribution. Simple kiging is performed with the transformed variable, then estimates and data are back-transformed to the original units. (Verly, 1983; Cressie, 1993; Emery, 2005, 2006)
- **Lognormal kriging.** The data are transformed into log-normal distribution. Simple kiging is performed with the transformed variable, then estimates and data are back-transformed to the original units. (Journel, 1980; Dowd, 1982; Cressie, 1993; Roth, 1998; Kishne, et al., 2003)

Although kriging is locally accurate, it is characterized by a smoothing effect in which small values are usually overestimated and large values underestimated; the variance of the kriging estimates is lower than the stationary domain variance. The smoothing effect is observed in kriging because it is a form of spatial regression. Due to the smoothing effect, kriged maps do not represent the spatial variability of the variable under study (Isaaks and Srivastava, 1989; Deutsch, 2002). Gesotatistical simulation corrects the smoothing of kriging.

2.6. Geostatistical Simulation

Gesostatistical simulation adds a random component to the kriging estimate to produce the proper variation at the model scale. Multiple equally-probable realizations of heterogeneity are obtained by random drawing different random components. These realizations provide an assessment of uncertainty about the properties of the variables being modeled (Journel, 1990).

Geostatistical simulation differs from kriging in two major aspects. Firstly, geostatistical simulation allows the calculation of the joint uncertainty when several locations are considered together. Secondly, in geostatistical simulation, reproduction of the global features and statistics (e.g., histogram and variogram) plays a more important role than local accuracy, which is the goal in kriging. (Deutsch and Journel, 1998). Note also that geostatistical simulation ensures reproduction of the spatial variability of the phenomenon under study in simulation between the simulated values and the data and between the simulated values and themselves.

Stochastic simulation is an important component of geostatistical analysis. It quantifies uncertainty in energy and mineral resources (Journel, 1974). Other applications consist of generating input for flow simulation (Deutsch, 2002) and calculating the likelihood of exceeding critical threshold in contamination studies (Kyriakidis and Journel, 2001).

There are many approaches that can be used for geostatistical simulation. The most popular and simplest simulation techniques for continuous variables are based on the assumption of multivariate Gaussianity. The multivariate Gaussian distribution is characterized by the property that all conditional and marginal distributions are also Gaussian. The simple kriging mean and variance are precisely the mean and variance of the local conditional distribution.

It is rare, however, that a geological variable is Gaussian; therefore, the data need to be transformed before analysis. Simulation is conducted in normal score or Gaussian units. When stochastic simulation is completed, the simulated values are back-transformed into original units.

There are many algorithms for Gaussian simulation. These include sequential Gaussian simulation, matrix approaches, moving average, turning bands and spectral methods. A brief description of two of the most popular, that is, sequential Gaussian simulation and matrix approach follows, please refer to David (1977), Journel and Huijbregts (1978) and Chiles and Delfiner (1999) for a review of the other approaches.

2.6.1. Matrix Simulation (LU Simulation)

The matrix simulation, also known as LU simulation, is based on the Cholesky decomposition of the positive definite covariance matrix (Alabert, 1987; Davis, 1987). The covariance matrix to be decomposed contains information on the covariance between the sample data, sample data and nodes to be simulated and between nodes to be simulated. It is of size $(n+N)\times(n+N)$, where n is the number of data and N is the number of nodes to be simulated, that is,

$$C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix},$$

where C_{11} is the data-to-data covariance, C_{12} is data-to-simulated nodes covariance, C_{21} is simulated nodes-to-data covariance and C_{22} is simulated nodes-to-simulated nodes covariance.

The Cholesky decomposition of the covariance matrix C results in the lower and upper triangular matrices L and U ($U = L^T$) such that

$$C = LU = \begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{bmatrix}. \tag{2.32}$$

A conditional simulated realization with LU approach can be generated as follows

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}, \tag{2.33}$$

or

$$\begin{aligned} y_1 &= L_{11} w_1, \\ y_2 &= L_{21}(L_{11}^{-1} y_1) + L_{22} w_2, \end{aligned} \tag{2.34}$$

where y_1 is an n by 1 column vector of normal scores of the conditioning data, y_2 is an N by 1 vector of simulated values, $w_1 = L_{11}^{-1} y_1$ and w_2 is an N by 1 vector of standard normal random values. Different standard normal random values in vector w_2 need to be generated in order to generate different realizations.

The simulation is fast and can be performed with high efficiency after the Cholesky decomposition is done. Cholesky decomposition needs to be performed only once, therefore this method can be particularly useful if the grid is relatively small and a large number of realizations is required.

2.6.2. Sequential Gaussian Simulation

Sequential Gaussian simulation is the most commonly adopted approach for large realizations. This approach is simple and flexible; it is based on decomposing the multivariate distribution into a series of conditional distributions for each location. Joint simulation of *N* variables is obtained by drawing sequentially from successive univariate conditional distributions (and from the marginal in the first draw) (Levy, 1937).

The aim of the sequential Gaussian simulation is to correct kriging by adding to the simple kriging estimate an independent normally distributed residual with mean of zero and variance equal to the simple kriging variance. The added residuals correct the smoothness of kriging and ensure that simulated values have the right variability. (Goovaerts, 1997).

The shape of the conditional distributions is Gaussian ensuring that the simulated realizations will be standard normally distributed. The variogram reproduction is ensured by using original data values and simulated nodes. The following is a short summary of the steps required for sequential Gaussian simulation (Deutsch, 2002):

1. Transform the variable to be standard normal.
2. Select a random path to visit each location to be simulated.
3. Visit each location one-by-one and perform simple kriging to find the mean and variance of the local conditional distributions.
4. Draw a random value from the Gaussian conditional distribution.
5. Consider the simulated value as a data for subsequent nodes.
6. Repeat steps 3-5 until all locations are simulated.

The use of random path is designed to avoid the artifacts in the simulated values. Multiple equally-probable realizations can be created by changing the random number seed, that is, changing the random path and drawing different values from the conditional distributions.

2.6.3. Gaussian Cosimulation

There are several algorithms available for cosimulation of multiple dependent variables in the multivariate Gaussian context. In particular, matrix simulation (LU) can be easily extended to account for multiple data types. This approach, however, is impractical due to the size restrictions on the covariance matrix (Deutsch, 2002).

The most popular approach for Gaussian cosimulation is based on the sequential Gaussian simulation. Local conditional distributions are calculated based on simple cokriging or collocated simple cokriging. Simple cokriging is based on the linear model of coregionalization to describe the spatial relationship between random variables, while simple collocated cokriging is based on the simpler Markov model I.

2.6.4. Review of Other Simulation Techniques

There are many other simulation techniques. Some do not rely on the assumption of the multivariate Gaussian distribution. The following are quite common.

- **Sequential indicator simulation**. Data are coded as indicators at a number of different thresholds. A variogram model is calculated for each threshold. Kriging of the indicator function is used for building the distribution of uncertainty. A simulated value is drawn from the resulting distribution (Journel and Isaaks, 1984; Chu, 1996; Emery, 2004).
- **P-field simulation.** P-field simulation is performed in two steps. Firstly, the local conditional distributions are obtained, then random field with a desired correlation structure is generated. At each location of the study domain, a value of the random variable is drawn using the known local conditional distributions (Froideveaux, 1993; Goovaerts, 1997; Saito and Goovaerts, 2002; Goovaerts, 2002).

2.7. Discussion

Geostatistics provides a set of statistical and mathematical tools for generating numerical models of regionalized variables that help in data processing and decision making.

Although many methods for geostatistical simulation and estimation have been proposed over the years; there are still a lot of longstanding problems associated with these techniques. These problems include string effect of kriging, variance inflation of collocated cokriging, development of combined measure of uncertainty, improving the multivariate simulation with correlated residuals and others.

The proposed work aims at investigating some of these issues and proposing novel theoretical and/or practical solutions for them.

CHAPTER 3

Correcting the String Effect

This chapter investigates the string effect of kriging. Two new approaches for correcting the kriging weights to avoid the assignment of large weights to data at the end of strings are proposed. Both proposed methods are shown to remove systematic bias in resource estimation.

Section 3.1 defines and illustrates the string effect of kriging and presents the motivation for a correction. **Section 3.2** reviews the methods currently applied to correct the string effect. **Section 3.3** and **Section 3.4** present two newly developed methods for resource estimation in a finite domain, that is, distance constrained kriging and finite domain kriging. The methods are explained in detail and illustrated with small examples. **Section 3.5** compares the two proposed kriging methods with a case study. A brief discussion of the results and the methods is presented in **Section 3.6**.

3.1. Introduction: String Effect

Data are collected along drillholes or wells. Finite strings of data are commonly encountered in natural resource applications in the following two situations: (1) strings of data are truncated by geologic or stratigraphic boundaries (Figure 3.1), and (2) strings of data are truncated by a local search ellipsoid (Figure 3.2).

When kriging is used with finite strings of data it can be frequently observed that outlying data in the strings receive higher weights (taking into account the sign) than all other data. This counter-intuitive weighting, referred to as the string effect, is illustrated in Figure 3.3 for one string of data (see also Deutsch, 1993). The three profiles of ordinary kriging weights are shown in Figure 3.3. These weights are "the string effect". They are obtained based on a spherical variogram model with a range of correlation equal to the distance from the estimation location to the string and with nugget effects of 0%, 25% and 75%, respectively. Figure 3.3 shows that the end samples in the string of data receive proportionally higher weights than the rest of the data in the string. Moreover, the difference in the weights increases with a decrease in the nugget effect. Note also that if the distance from the string to the estimation location increases then the structure of ordinary kriging weights remains unchanged. The only situation in which border samples do not receive significantly higher weights is when the estimation location is close to the string, see Figure 3.4. The redundancy of the data that causes the string effect is less important when the estimation location is close to the string.

The unusual pattern of kriging weights is theoretically valid and is caused by the egrodicity assumption of kriging. The outlying samples are given a large weight because the data-to-data (left hand side) kriging matrix views such samples as less redundant than the rest of the samples (Deutsch, 1994). The outlying samples are informative of the half space beyond the string of data, thus, they receive higher weights.

In the case of a bounded geologic domain or horizon, there is no volume outside the finite study area that would justify large weighting of the boundary samples in the strings. In the case of a limited search neighborhood, there are data outside of finite strings, but these data may not be reflected by the end points of the search neighborhood. Thus, there may be no justification for the string effect. The overweighting may have an especially strong impact when the domain under study is non-stationary, that is, when the data at the ends of the finite string the data values are higher or lower than in the middle of the string. These problems motivate development of methods for correcting the string effect.

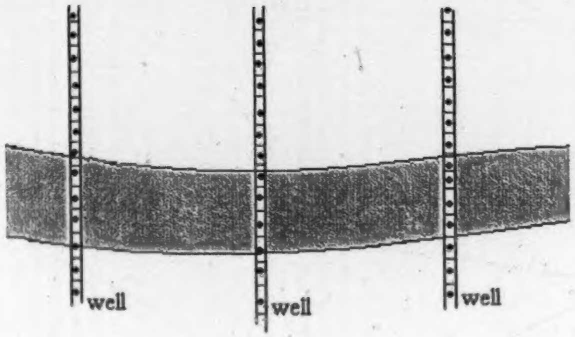

well
well
well

Figure 3.1: Finite strings of data as a result of stratigraphic boundary truncation (redrawn from Deutsch, 1993).

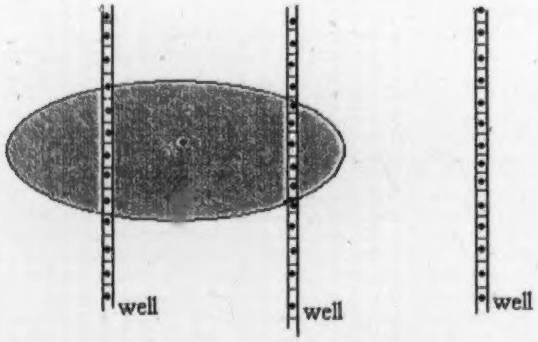

well
well
well

Figure 3.2: Finite strings of data as a result of a limited search (redrawn from Deutsch, 1993).

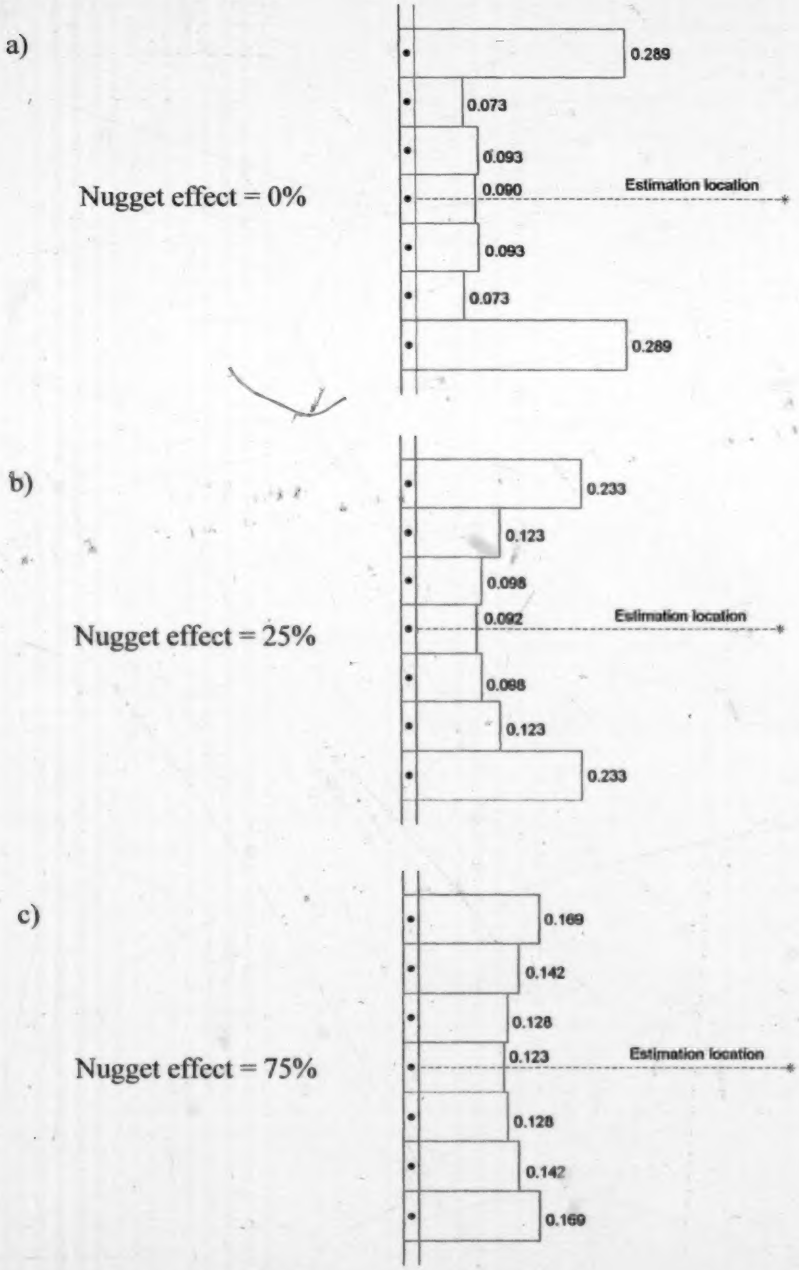


Figure 3.3: Profiles of ordinary kriging weights obtained for the estimation point located on a distance equal to the range of correlation from the string. Results are calculated based on a spherical variogram model with a nugget effect of 0% (a), 25% (b) and 75% (c) using string of 7 data.

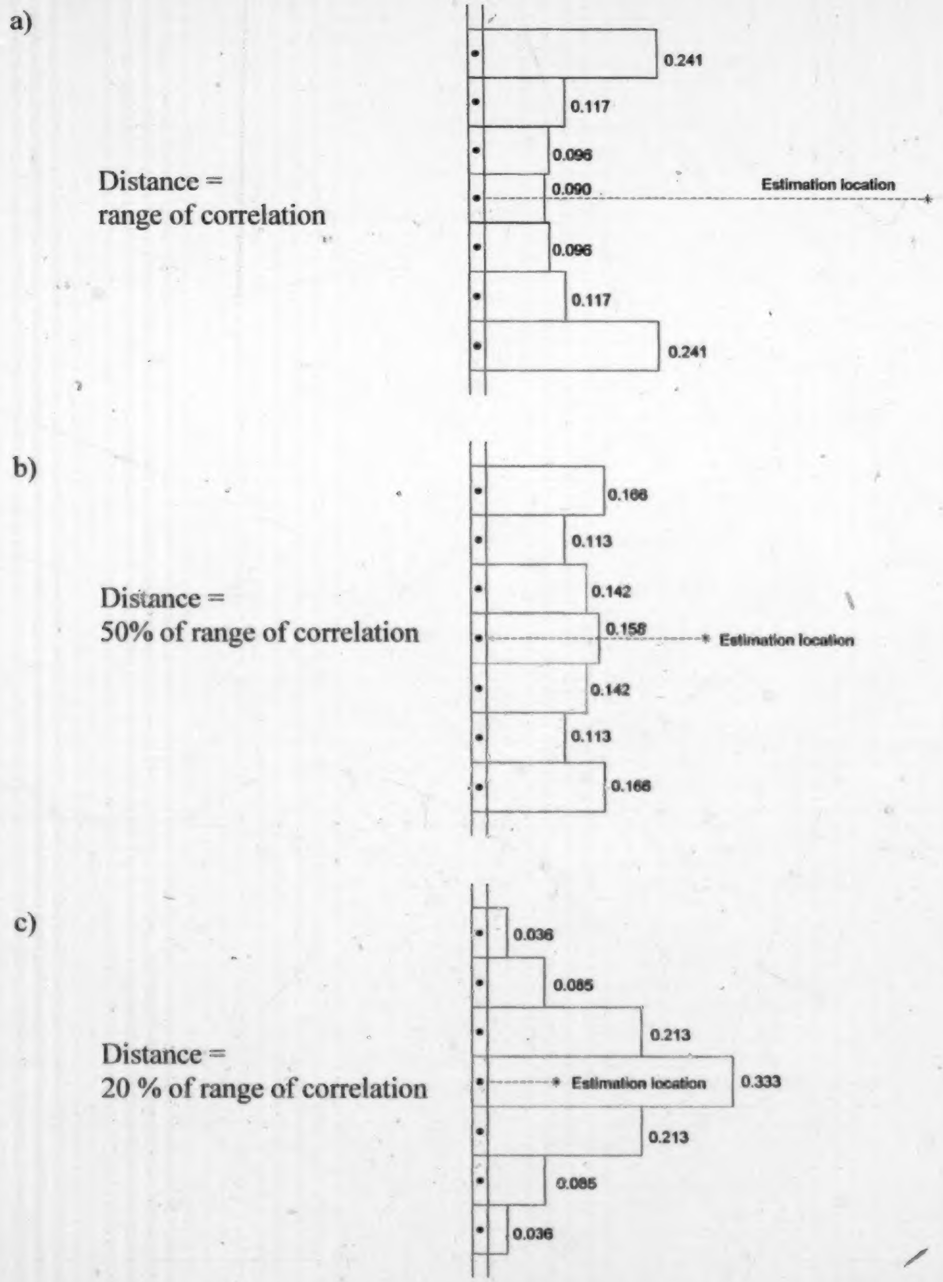


Figure 3.4: Profiles of the ordinary kriging weights obtained for the estimation point located on the distance equal to the range of correlation (a), 50% of the range of correlation (b) and 20% of the range of correlation (c). Results are calculated based on a spherical variogram model with a nugget effect of 20% using a string of 7 data.

3.2. Methods for Fixing the String Effect

What follows is a short description of five empirical solutions (Deutsch, 1993; Deutsch, 1994; and Saito et al. 2005) to the string effect. All of these solutions are ad hoc; they change either the data configuration or modify the covariance function in order to ensure that all samples within the same string have the same average redundancy. They are not fully automatic with a well defined measure of optimality.

3.2.1. Quick Fix

A quick fix approach to the string effect restricts kriging to only use two samples from every string of data. In this way, both of the samples are equally informative and redundant. This approach for correcting the string effect does not take into account all available information on the variable of interest. Therefore, a major disadvantage of this quick fix approach is that too few data are used in estimation which may result in unreliable estimates.

3.2.2. Extend the String

Artificial or phantom data are added to each end of the string. Then, kriging is performed to find the weights for all data in the extended string. The weights assigned to the phantom data are set to zero and the remaining weights rescaled to sum to one. This approach can not be applied with simple kriging, since simple kriging weights do not sum to one. Other serious limitations include: (1) the data configuration may contain multiple strings with different numbers of samples in them. Thus, adding phantom data with the same support to the ends of each string can result in unfair weighting of strings of different length. (2) The data close to the phantom data could also receive significant weights. Thus, the edge effect may not be removed completely. Moreover, in order to correctly add the phantom data to each end of the string, the local direction of the string must be known. This may create difficulties in practical implementation.

3.2.3. Use Simple Kriging

The string effect is not as pronounced when estimating with simple kriging, see Figures 3.5-3.6 and Deutsch (1993). Simple kriging requires the mean to be specified. The mean is first calculated by declustering or ordinary kriging. Then, simple kriging is applied to find the value of the variable of interest at the estimation location.

The correction with simple kriging is not complete: weights assigned to a string of data after performing the above described steps still reveal artifact large weighting of the end data in the string.

3.2.4. Wrap the String

The wrap the string method deals with the assumption of an infinite domain by modifying the data-to-data covariance in the left-hand side kriging covariance matrix. Specifically, for n data aligned in a finite string each separated from its neighbor conditioning data by a distance vector **h**, the covariance between any i-th data Z_i and j-th Z_j in the string is modified to become

$$Cov(Z_i, Z_j) = C(k\mathbf{h}), \tag{3.1}$$

where

$$k = \min\{(j - i + n), (i - j)\}.$$

The modified left-hand side covariance matrix corresponds to correcting the covariance between the data to show all data as equally relevant/redundant with all others. The wrap the string method is also applied in other image processing and geophysical applications, where the infinite domain assumption also causes problems in estimation (e.g., Aki and Richards, 1980).

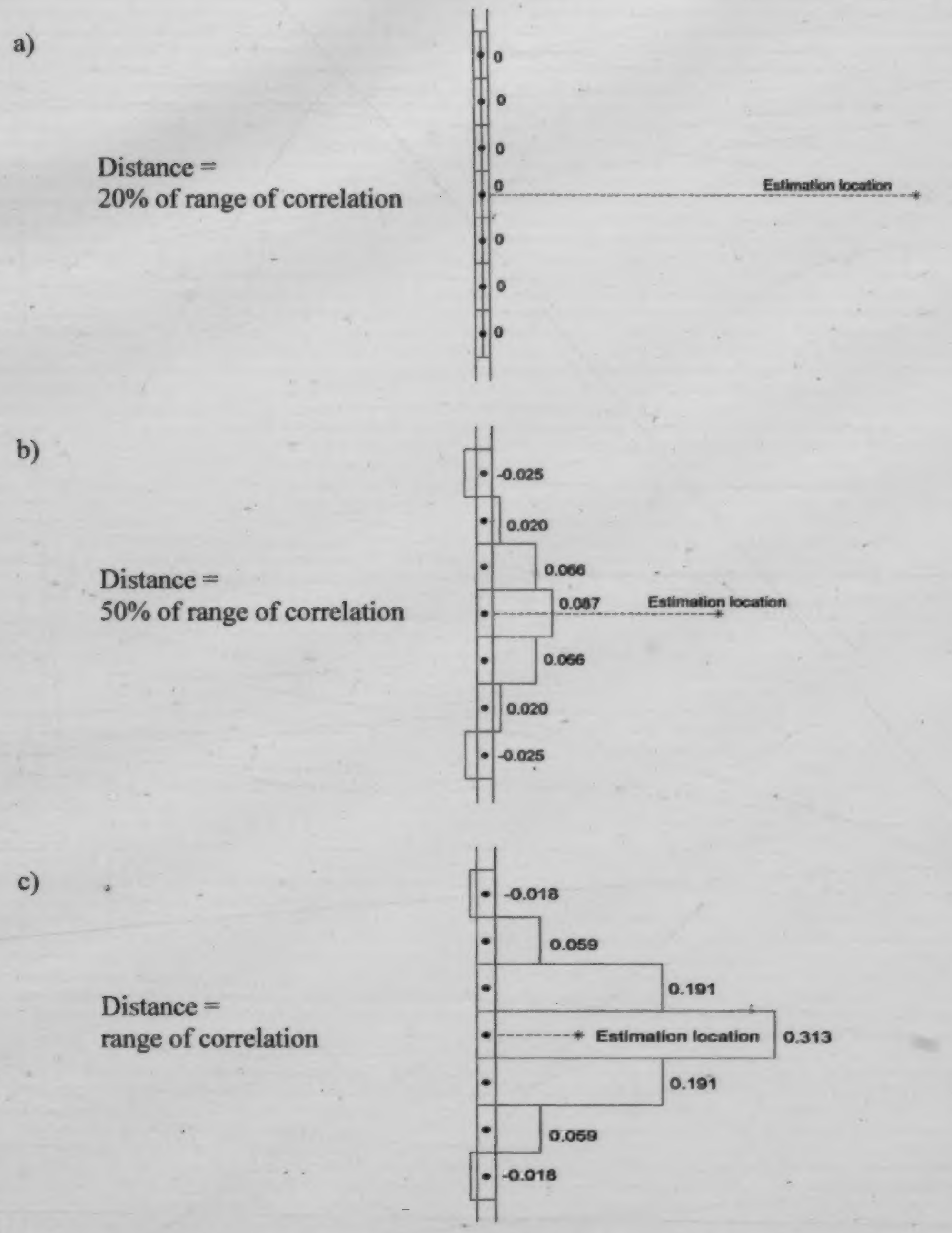

a)
Distance =
20% of range of correlation
0
0
0
0
Estimation location
0
0
0
b)
Distance =
50% of range of correlation
-0.025
0.020
0.066
0.087
Estimation location
0.066
0.020
-0.025
c)
Distance =
range of correlation
-0.018
0.059
0.191
Estimation location
0.313
0.191
0.059
-0.018

Figure 3.5: Profiles of the simple kriging weights obtained for the estimation point located on the distance equal to to the range of correlation (a), 50% of the range of correlation (b) and 20% of the range of correlation (c). Results are calculated based on a spherical variogram model with a nugget effect of 20% using a string of 7 data..

3.2.5. Finite Domain Kriging of Deutsch (1994)

Finite domain kriging approach can be formulated as follows. In the case of resource estimation using a single string of data, the finite domain kriging estimate of the resource of interest at location $\mathbf{u}$ is found as a linear combination of neighboring observations $Z(\mathbf{u}_i), i = 1,\ldots,n(\mathbf{u})$,

$$Z^*_{FDK}(\mathbf{u}) = \sum_{i=1}^{n(\mathbf{u})} \nu_i(\mathbf{u}) Z(\mathbf{u}_i), \tag{3.2}$$

where $\mathbf{v}^T = (\nu_1(\mathbf{u}),\ldots,\nu_{n(\mathbf{u})}(\mathbf{u}))^T$ denotes the vector of the finite domain kriging weights calculated from the following system of equations for the estimation location $\mathbf{u}$,

$$\begin{gathered} \sum_{i=1}^{n(\mathbf{u})} \nu_i(\mathbf{u}) r(\mathbf{u}_i,\mathbf{u}_j) + \mu_{FDK}(\mathbf{u}) = Cov(Z(\mathbf{u}),Z(\mathbf{u}_j)), \quad j = 1,\cdots,n(\mathbf{u}), \\ \sum_{i=1}^{n(\mathbf{u})} \nu_i(\mathbf{u}) = 1, \end{gathered} \tag{3.3}$$

where $r(\mathbf{u}_i,\mathbf{u}_j), \quad i,j = 1,\ldots,n(\mathbf{u})$, denotes the redundancy measure given by

$$r(\mathbf{u}_i,\mathbf{u}_j) = Cov(Z(\mathbf{u}_i),Z(\mathbf{u}_j)) - \left[\overline{C}(\mathbf{u}_j) - \overline{C}(\mathbf{u}_i)\right], \tag{3.4}$$

where $\overline{C}(\mathbf{u}_i)$ is the redundancy term of sample $Z(\mathbf{u}_i)$ of size N is given by

$$\overline{C}(\mathbf{u}_i) = \frac{1}{N}\sum_{j=1}^{N} Cov(Z(\mathbf{u}_i),Z(\mathbf{u}_j)); \tag{3.5}$$

$\mu_{FDK}(\mathbf{u})$ is the Lagrange parameter and, as before, $Cov(Z(\mathbf{u}),Z(\mathbf{u}_j)), \quad j = 1,\cdots,n(\mathbf{u})$, is the data-to-estimation point covariance function.

The difference with finite domain kriging (FDK) lies in the left-hand data-to-data covariance matrix of kriging. The covariance function in the left-hand data-to-data covariance matrix is replaced by the redundancy measure $r(\mathbf{u}_i,\mathbf{u}_j)$. This measure ensures that all samples within the same string (N) have exactly the same average redundancy. In particular, the redundancy correction (3.4) results in increasing the redundancy of outlying samples, and decreasing that of central samples.

One of the most substantial drawbacks of the finite domain kriging estimator is that it, on the contrary to all other kriging estimators, does not possess the exactitude property by which all data values are honored at their exact locations. Although kriging estimates are not usually calculated at existing data locations (except for cross validation as a measure of goodness of kriging results), the lack of the exactitude property can create strong discontinuities in the estimates next to data locations.

The estimation variance of the finite domain kriging estimator is:

$$\sigma^2_{FDK} = \sigma^2 - \sum_{i=1}^{n(\mathbf{u})} \nu_i(\mathbf{u}) Cov(Z(\mathbf{u}), Z(\mathbf{u}_i)) - \mu_{FDK}(\mathbf{u}) \geq 0.$$

Due to the lack of the exactitude property the estimation variance at the data locations may not be zero.

In the case of finite domain estimation based on L strings of data, finite domain kriging estimates are first found for the estimation location separately based on $n_l, l = 1,\ldots,L,$ data from each string

$$Z^{*l}_{FDK}(\mathbf{u}) = \sum_{i=1}^{n_l(\mathbf{u})} \nu_i^l(\mathbf{u}) Z(\mathbf{u}_i^l), \tag{3.6}$$

where n_l locations $\mathbf{u}_i^l$ are used from string l to provide an estimate at location $\mathbf{u}$ and $\mathbf{v}^T = (\nu_1(\mathbf{u}),\ldots,\nu_{n(\mathbf{u})}(\mathbf{u}))^T$ denotes the vector of the finite domain kriging weighs calculated for string l.

Using the finite domain kriging estimates calculated for the location of interest $\mathbf{u}$, the final finite domain kriging estimate for this location is obtained as

$$Z^{**}_{FDK} = \sum_{l=1}^{L} \omega_l(\mathbf{u}) Z^{*l}_{FDK} \tag{3.7}$$

based on the ordinary kriging system of equations

$$\begin{gathered} \sum_{k=1}^{L} \omega_k(\mathbf{u}) \overline{C}(l-k) + \mu(\mathbf{u}) = \overline{C}(l-\mathbf{u}), \quad l = 1,\cdots,L, \\ \sum_{k=1}^{L} \omega_k(\mathbf{u}) = 1, \end{gathered} \tag{3.8}$$

where the string-to-string covariance function $\overline{C}(l-k)$ and string-to-unknown covariance function $\overline{C}(l-\mathbf{u})$ are given, respectively, by

$$\overline{C}(l-k) = \frac{1}{n_l n_k} \sum_{i=1}^{n_l} \sum_{j=1}^{n_k} Cov(Z(\mathbf{u}_i^l), Z(\mathbf{u}_j^k)) \tag{3.9}$$

$$\text{and} \quad \overline{C}(l-\mathbf{u}) = \frac{1}{n_l}\sum_{i=1}^{n_l} C(Z(\mathbf{u}_i^l), Z(\mathbf{u})). \tag{3.10}$$

Note that due to the fact that each of the finite domain kriging estimates of the unknown resource value at location **u** is not exact, the final finite domain kriging estimate is also not exact.

In conclusion, the finite domain kriging proposed by Deutsch (1994) never gained popularity due to complexity, lack of optimality and absence of exactitude property.

3.3. Distance Constrained Kriging

In order to correct the string effect the weights are constrained to have a certain reasonable influence structure. Specifically, the weights assigned to the string of data are ordered with respect to distance from the unsampled location: the closest data in the string is constrained to receive the largest weight; the second closest data is constrained to receive the second largest weight and so on. The data in the string located furthest from the estimation location is assigned the smallest weight.

3.3.1. Methodology

Let us consider n adjacent data at locations $\mathbf{u}_i, i=1,\ldots,n,$ aligned in a string. Consider the problem of estimating the value of a variable of interest Z at an unsampled location $\mathbf{u}_0$ using the distance constrained kriging approach.

The distance constrained simple kriging (DCSK) estimator provides a model of the unsampled value $Z(\mathbf{u}_0)$ as the following linear combination of the data in a string $Z(\mathbf{u}_i), i=1,\ldots,n,$ and the stationary data mean m

$$Z^*_{DCSK}(\mathbf{u}_0) = \sum_{i=1}^{n} \lambda_{DCSK,i}(\mathbf{u}_0) Z(\mathbf{u}_i) + \left[1 - \sum_{i=1}^{n} \lambda_{DCSK,i}(\mathbf{u}_0)\right] m, \tag{3.11}$$

where $\lambda_{DCSK,i}(\mathbf{u}_0) = \lambda_{DCSK,i}$ denotes the DCSK weight of the i-th sample found by minimizing the estimation variance σ^2_{est}

$$\min_{\lambda_{DCSK}} \sigma^2_{est} = \sigma^2 - 2\sum_{i=1}^{n} \lambda_{DCSK,i} Cov(Z(\mathbf{u}_i), Z(\mathbf{u}_0)) + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_{DCSK,i}\lambda_{DCSK,j} Cov(Z(\mathbf{u}_i), Z(\mathbf{u}_j)) \quad (3.12)$$

subject to

$$\lambda_{DCSK,i} > \lambda_{DCSK,j}, \qquad \text{if } d_i < d_j, \qquad \text{for each } i, j = 1, \ldots, n, \quad (3.13)$$

where σ^2 is the stationary variance; $Cov(Z(\mathbf{u}_i), Z(\mathbf{u}_j))$ and $Cov(Z(\mathbf{u}_i), Z(\mathbf{u}_0))$ are the data-to-data covariance and the data-to-estimation point covariance, respectively; and d_i denotes the distance from the estimation location to the i-th data point in the string, $i, j = 1, \ldots, n$.

Distance constrained ordinary kriging (DCOK) estimates the value at the location of interest $\mathbf{u}_0$ as

$$Z^*_{DCOK}(\mathbf{u}_0) = \sum_{i=1}^{n} \lambda_{DCOK,i}(\mathbf{u}_0) Z(\mathbf{u}_i), \quad (3.14)$$

where $\lambda_{DCOK,i}(\mathbf{u}_0) = \lambda_{DCOK,i}$, denotes the DCOK weight of the i-th sample found by minimizing the estimation variance σ^2_{est}

$$\min_{\lambda_{DCOK}} \sigma^2_{est} = \sigma^2 - 2\sum_{i=1}^{n} \lambda_{DCOK,i} Cov(Z(\mathbf{u}_i), Z(\mathbf{u}_0)) + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_{DCOK,i}\lambda_{DCOK,j} Cov(Z(\mathbf{u}_i), Z(\mathbf{u}_j)) \quad (3.15)$$

subject to

$$\lambda_{DCOK,i} > \lambda_{DCOK,j}, \qquad \text{if } d_i < d_j, \qquad \text{for each } i, j = 1, \ldots, n, \quad (3.16)$$

and

$$\sum_{i=1}^{n} \lambda_{DCOK,i} = 1. \quad (3.17)$$

3.3.2. Implementation

The distance constrained simple kriging and distance constrained ordinary kriging formulations given in Equations (3.11)-(3.13) and (3.14)-(3.17), respectively, can be solved by using a non-linear constrained optimization. However, they can also be reformulated in such way that a simple nonlinear unconstrained optimization can be applied for solving them.

Let us consider distance constrained simple kriging first. The minimization problem stated in Equations (3.12)-(3.13) is equivalent to the following

$$\min_{\xi} \sigma_{est}^2 = \sigma^2 - 2\sum_{i=1}^{n} \lambda_{DCSK,i}(\xi) Cov(X_i, X(\mathbf{u})) + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_{DCSK,i}(\xi)\lambda_{DCSK,j}(\xi) Cov(X_i, X_j) \tag{3.18}$$

where ξ is a new parameter vector of size n by 1 with respect to which minimization is performed; the DCSK weights $\lambda_{DCSK,i}(\xi)$, $i = 1,\ldots,n$, are as follows

$$\lambda_{DCSK,i}(\xi) = [n - k_i + 1]\text{ - th largest element in vector } \xi\,, \tag{3.19}$$

where k_i denotes the rank of the distance from the location where we are estimating to the i-th data in the string; d_i, is a vector of all distances $d = [d_1\ d_2 \ldots d_n]$ sorted in ascending order.

Similarly, we can rewrite the minimization problem stated in Equations (3.15)-(3.17) for the distance constrained ordinary kriging as follows

$$\min_{\xi} \sigma_{est}^2 = \sigma^2 - 2\sum_{i=1}^{n} \lambda_{DCOK,i}(\xi) Cov(X_i, X(\mathbf{u})) + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_{DCOK,i}(\xi)\lambda_{DCOK,j}(\xi) Cov(X_i, X_j) \tag{3.20}$$

where ξ is a new parameter vector of size n by 1 with respect to which minimization is performed; the DCOK weights $\lambda_{DCOK,i}(\xi)$, $i = 1,\ldots,n$, are given by

$$\lambda_{DCOK,i}(\xi) = \frac{[n - k_i + 1]\text{ - th largest element in vector } \xi}{\text{sum of all element in } \xi}. \tag{3.21}$$

3.3.2.1. Program

For solving the distance constrained simple kriging minimization problem (3.18)-(3.19) and the distance constrained ordinary kriging minimization problem (3.20)-(3.21), the optimization subroutine MINF1 from the Scientific Subroutine Library II (SSL II) can be used. This subroutine is designed to perform minimization of a function with several variables using the revised quasi-Newton method based on function values only. For convenience this subroutine is incorporated into GSLIB program *kt3d* (Deutsch and Journel, 1998).

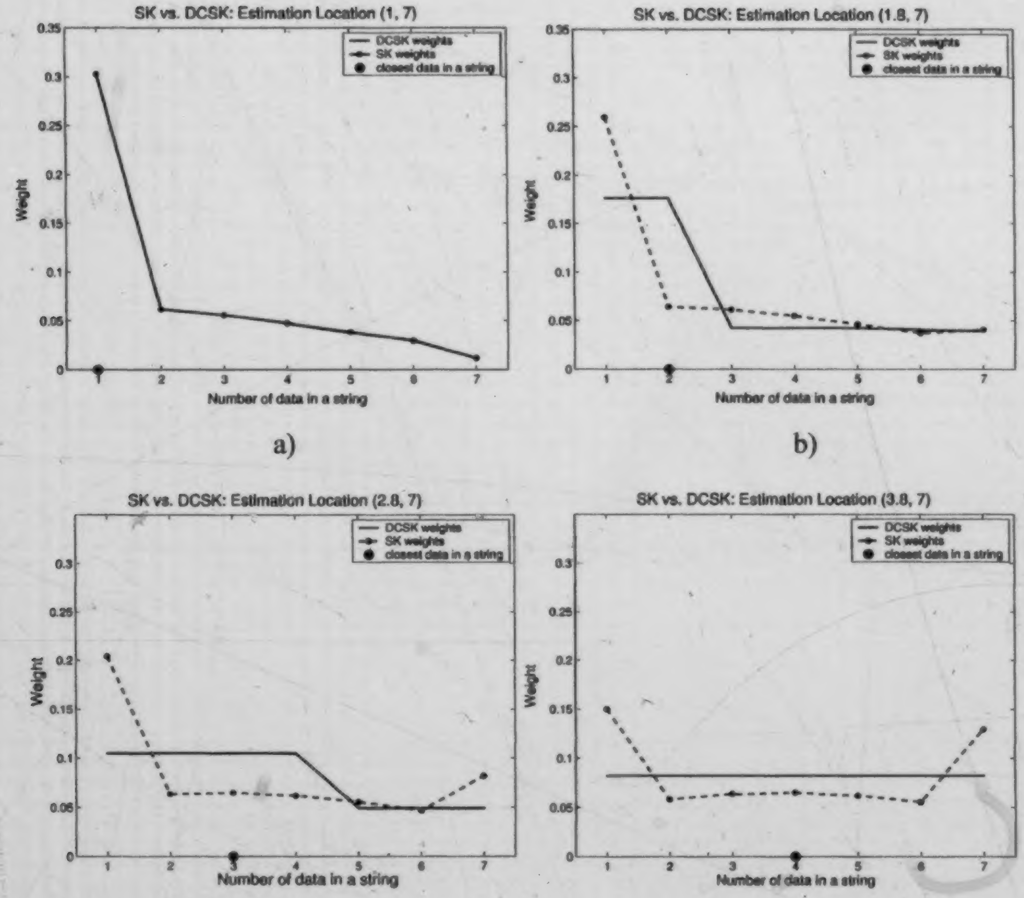
It is worth noting, that minimization problems (318) and (3.19) results in several local optimums for ξ (n! to be specific). All these omptimums are permutations of the same vector. Therefore, all these solutions correspond to the same unique vector $\lambda_{DCOK}(\xi)$ obtained by transformation (3.19) (for simple kriging) or (3.20) for ordinary kriging.

Also note that due to minimization performed in distance constrained kriging program for each location of the study domain, the time required to complete distance contrained kriging estimation increases substantially campared to tradional kriging. As a starting point in optimization for distance constrained kriging weights, traditional (ordinary or simple) kriging weights can be taken.

3.3.3. Small Examples

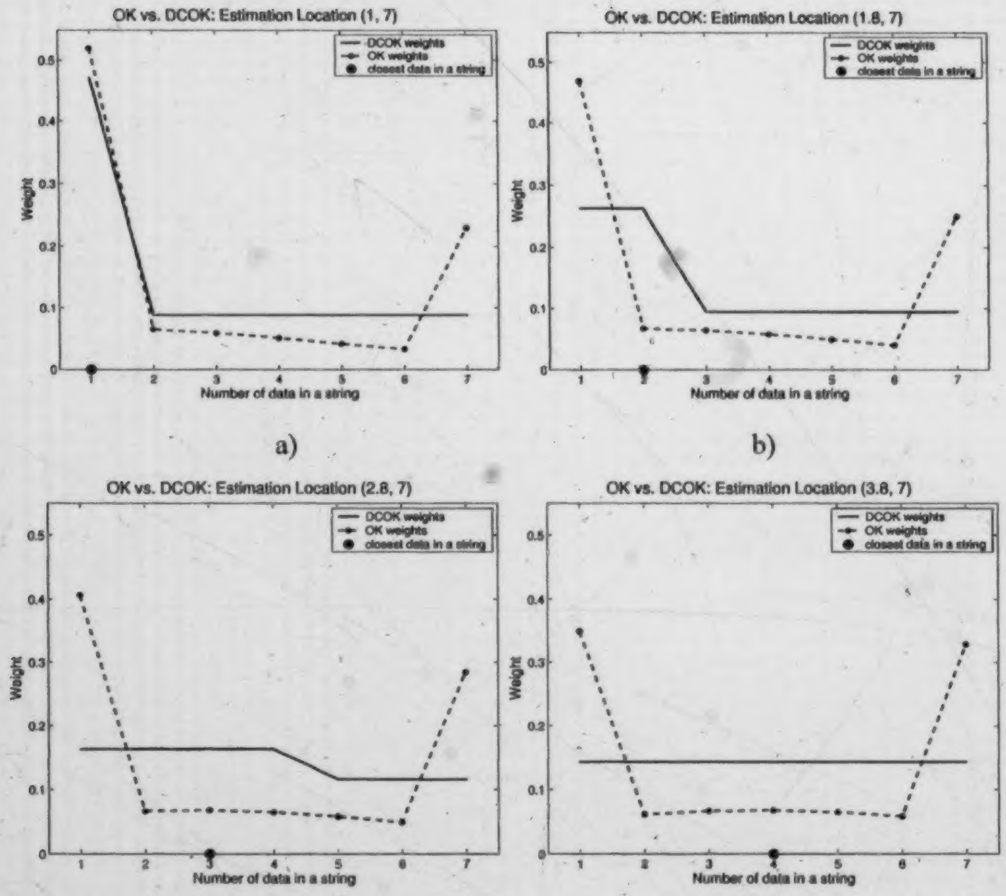
Several small studies compare the kriging weights obtained using the traditional simple and ordinary kriging approaches with distance constrained kriging. The weights were calculated for four estimation locations, (1, 7), (1.8, 7), (2.8, 7) and (3.8,7), based on the string of 7 data located at (1,0), (2,0), (3,0), (4,0), (5,0), (6,0) and (7,0). An isotropic spherical variogram with a sill of one and range of 20 is considered. This 1-D example was chosen small enough to easily visualize the results yet large enough to show realistic variations in the results. The conclusions drawn from this example are considered reasonably general.

The weights are shown in Figure 3.6 for simple kriging and distance constrained simple kriging. Figure 3.7 shows ordinary kriging and distance constrained ordinary kriging. Note how the distance constrained kriging approaches reduce the artificially high weights given to the end samples of the string.



a) b) c) d)

Figure 3.6: Profiles of the distance constrained simple kriging (solid line) and simple kriging (dashed line) weights. Results shown for estimation locations: a) (1, 7); b) (1.8, 7); c) (2.8, 7); and d) (3.8, 7) are obtained using a spherical variogram model with a range of 20 based on a string of 7 data located at (1,0), (2,0), (3,0), (4,0), (5,0), (6,0) and (7,0), respectively. The closest data in a string is denoted by a dark circle.



a) b)

c) d)

Figure 3.7: Profiles of the distance constrained ordinary kriging (solid line) and ordinary kriging (dashed line) weights. Results shown for estimation locations: a) (1, 7); b) (1.8, 7); c) (2.8, 7); and d) (3.8, 7) are obtained using a spherical variogram model with a range of 20 based on a string of 7 data located at (1,0), (2,0), (3,0), (4,0), (5,0), (6,0) and (7,0), respectively. The closest data in a string is denoted by a dark circle.

3.3.4. Properties

As with the traditional simple and ordinary kriging techniques, the distance constrained kriging approaches have the following characteristics:

1 Distance constrained kriging estimators are unbiased.

2 Distance constrained kriging estimators are exact interpolators, that is, they reproduce conditioning data at their locations.

3 Distance constrained kriging provides estimates for the unsampled value of the variable of interest according to its spatial continuity described by the covariance function. Distance constrained kriging takes into account the redundancy of data in the string and closeness of the data in the string to an estimation location.

The distance constrained kriging estimate are obtained as a linear combination of data in string that minimizes estimation variance; however, in the distance constrained kriging approach the estimation variance is minimized according to the distance constraints, thus, the estimate produced is characterized by the same or higher estimation variance than respective kriging estimates, that is,

$$\sigma_{est}^{DCK} \geq \sigma_{est}^{K}, \tag{3.22}$$

where σ_{est}^{K} and σ_{est}^{DCK} are the estimation variance in the traditional kriging and in the distance constrained kriging.

Note also that distance constrained kriging approaches are also characterized by the following property. The weights in the distance constrained kriging are assigned to the data in the string sorted according to the distance from these data to the estimation location.

Thus, the distance from the data to the estimation location has an affect on the resulting estimate; however, unlike the inverse distance technique, this affect is additionally corrected by the spatial continuity of the variable under study.

3.3.5. Distance Constrained Kriging: Generalization to the Case of Multiple Strings

Let us consider K strings of n_k data each at locations $\mathbf{u}_i^k, k = 1,\ldots,K, i = 1,\ldots,n_k$. Then the value of the variable of interest Z at an unsampled location $\mathbf{u}_0$ in the distance constrained simple kriging approach is given by the following linear combination of the data in strings $Z(\mathbf{u}_i^k)$ and the stationary domain mean m

?

$$Z^*_{DCSK}(\mathbf{u}_0) = \sum_{k=1}^{K}\sum_{i=1}^{n_k} \lambda_{DCSK,i}^k(\mathbf{u}_0) Z(\mathbf{u}_i^k) + \left[1 - \sum_{k=1}^{K}\sum_{i=1}^{n_k} \lambda_{DCSK,i}^k(\mathbf{u}_0)\right] m, \qquad (3.23)$$

where $\lambda_{DCSK,i}^k(\mathbf{u}_0) = \lambda_{DCSK,i}^k$ denotes the DCSK weight of the i-th sample in the k-th string, $i = 1,\ldots,n_k$, $k = 1,\ldots,K$, found by minimizing the estimation variance σ_{est}^2

$$\min_{\lambda_{DCSK}} \sigma_{est}^2 = \sigma^2 - 2\sum_{k=1}^{K}\sum_{i=1}^{n_k} \lambda_{DCSK,i}^k(\mathbf{u}_0) Cov(Z(\mathbf{u}_i^k), Z(\mathbf{u}_0)) + \sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{l=1}^{K}\sum_{j=1}^{n_l} \lambda_{DCSK,i}^k(\mathbf{u}_0)\lambda_{DCSK,j}^l(\mathbf{u}_0) Cov(Z(\mathbf{u}_i^k), Z(\mathbf{u}_j^l)) \qquad (3.24)$$

subject to

$$\lambda_{DCSK,i}^k > \lambda_{DCSK,j}^l, \quad \text{if } d_i^k < d_j^l, \quad \text{for each } k,l = 1,\ldots,K, i = 1,\ldots,n_k, j = 1,\ldots,n_l, \qquad (3.25)$$

where σ^2 is the stationary variance; $Cov(Z(\mathbf{u}_i^k), Z(\mathbf{u}_j^k))$ and $Cov(Z(\mathbf{u}_i^k), Z(\mathbf{u}_0))$ stands for the data-to-data covariance and the data-to-estimation point covariance, respectively, and d_i^k denotes the distance from the estimation location to the i-th data point in the k-th string, $k,l = 1,\ldots,K, i = 1,\ldots,n_k, j = 1,\ldots,n_l$.

The distance constrained ordinary kriging (DCOK) estimates the value at the location of interest $\mathbf{u}_0$ as

$$Z^*_{DCOK}(\mathbf{u}_0) = \sum_{k=1}^{K}\sum_{i=1}^{n_k} \lambda_{DCOK,i}^k(\mathbf{u}_0) Z(\mathbf{u}_i^k), \qquad (3.26)$$

where $\lambda_{DCOK,i}^k(\mathbf{u}_0) = \lambda_{DCOK,i}^k$ denotes the DCOK weight of the i-th sample in the k-th string, $i = 1,\ldots,n_k$, $k = 1,\ldots,K$, found by minimizing the estimation variance σ_{est}^2

$$\min_{\lambda_{DCOK}} \sigma_{est}^2 = \sigma^2 - 2\sum_{k=1}^{K}\sum_{i=1}^{n_k} \lambda_{DCOK,i}^k(\mathbf{u}_0) Cov(Z(\mathbf{u}_i^k), Z(\mathbf{u}_0)) + \sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{l=1}^{K}\sum_{j=1}^{n_l} \lambda_{DCOK,i}^k(\mathbf{u}_0)\lambda_{DCOK,j}^l(\mathbf{u}_0) Cov(Z(\mathbf{u}_i^k), Z(\mathbf{u}_j^l))$$

(3.27)

subject to

$$\lambda^{k}_{DCOK,i} > \lambda^{l}_{DCOK,j}, \quad \text{if } d^{k}_{i} < d^{l}_{j}, \quad \text{for each } \ k,l = 1,\ldots,K,\ i = 1,\ldots,n_k,\ j = 1,\ldots,n_l, \tag{3.28}$$

$$\text{and} \quad \sum_{k=1}^{K}\sum_{i=1}^{n_k} \lambda^{k}_{DCOK,i} = 1, \tag{3.29}$$

Note that the four properties for the distance constrained kriging estimators outlined for the single string case also hold for the case of multiple strings.

3.4. Finite Domain Kriging

While the distance constrained kriging approach, contrary to the many ad-hoc solutions (Section 3.2), fixes the string effect without changing the data configuration or the covariance function, it introduces many constraints that may lead to suboptimal estimation. Other theoretically sound alternatives for correcting the string effect are warranted. An alternative, called finite domain kriging, is developed below.

The traditional (simple or ordinary) kriging technique is performed as many times as there are conditioning data, that is, *n* times. Each time kriging is based on the *k* closest data from the string. Each kriging is optimal, yet with different smoothing and a different treatment of data at the end of strings. The kriging weights to be used for finite domain estimation are the average of the weights from the *n* successive kriging runs. These weights do not give undue influence to data values at the end of strings.

3.4.1. Methodology

Let us consider *n* adjacent data $i = 1,\ldots,n,$ at locations $\mathbf{u}_i, i = 1,\ldots,n,$ aligned in a string. Consider now the problem of estimating the value of a variable of interest Z at an unsampled location $\mathbf{u}_0$ using the finite domain kriging approach.

The finite domain simple kriging (FDSK) provides a model of the unsampled value $Z(\mathbf{u}_0)$ at location $\mathbf{u}_0$ as the following linear combination of the data in a string $Z(\mathbf{u}_i)$, $i=1,\dots,n$, and the stationary domain mean m

$$Z_{FDSK}*(\mathbf{u}_0)=\frac{1}{n}\sum_{k=1}^{n}\left[\sum_{i=1}^{k}\lambda^k_{SK,i}(\mathbf{u}_0)Z_i^k+\left(1-\sum_{i=1}^{k}\lambda^k_{SK,i}(\mathbf{u}_0)\right)m\right] \tag{3.30}$$

where $Z^k=(Z_1^k,\dots,Z_k^k)$, $k=1,\dots,n$, denotes the vector of k closest data in a string to the estimation location $\mathbf{u}_0$; $\boldsymbol{\lambda}_{SK}^{k,T}=(\lambda^k_{SK,1},\dots,\lambda^k_{SK,k})^T$, $k=1,\dots,n$, denotes the vector of the simple kriging weights calculated from the normal system of equations for the estimation location $\mathbf{u}_0$ based on the k closest data in the string

$$\sum_{i=1}^{k}\lambda^k_{SK,i}Cov(Z_j^k,Z_i^k)=Cov(Z(\mathbf{u}_0),Z_i^k),\quad j=1,\cdots,k, \tag{3.31}$$

where $Cov(Z_j^k,Z_i^k)$ and $Cov(Z(\mathbf{u}_0),Z_i^k)$ denote data-to-data covariance and data-to-estimation point covariance, respectively, $i,j=1,\cdots,k$, $k=1,\dots,n$.

Note that the finite domain kriging estimator defined in Equation (3.30) can be rewritten in the following simpler form

$$Z_{FDSK}*(\mathbf{u}_0)=\sum_{i=1}^{n}\lambda_{FDSK,i}(\mathbf{u}_0)Z_i(\mathbf{u}_0)+\left(1-\sum_{i=1}^{n}\lambda_{FDSK,i}(\mathbf{u}_0)\right)m, \tag{3.32}$$

where $\boldsymbol{\lambda}^T_{FDSK}(\mathbf{u}_0)=(\lambda_{FDSK,1}(\mathbf{u}_0),\dots,\lambda_{FDSK,n}(\mathbf{u}_0))^T=(\lambda_{FDSK,1},\dots,\lambda_{FDSK,n})^T$ denotes the vector of the finite domain simple kriging weights calculated as

$$\lambda_{FDSK,i}=\frac{1}{n}\sum_{k=1}^{n}\tilde{\lambda}^k_{SK,i}, \tag{3.33}$$

where

$$\tilde{\lambda}^k_{SK,i}=\begin{cases}\lambda^k_{SK,m}, & \text{if } i\text{-th data is within } k \text{ closest data in the string to the estimation location } \mathbf{u}_0;\\ & \text{index } m \text{ denotes the order of data } Z_i(\mathbf{u}_0) \text{ in the vector } Z^k;\\ 0, & \text{otherwise.}\end{cases}$$

The estimation variance at the estimation location $\mathbf{u}_0$ for the finite domain simple kriging can be easily calculated as follows

$$\sigma^2_{FDSK}(\mathbf{u}_0) = \sigma^2 - 2\sum_{i=1}^{n} \lambda_{FDSK,i} Cov(Z(\mathbf{u}_i), Z(\mathbf{u}_0)) + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_{FDSK,i}\lambda_{FDSK,j} Cov(Z(\mathbf{u}_i), Z(\mathbf{u}_j)) \tag{3.34}$$

The finite domain ordinary kriging (FDOK) provides a model of the unsampled value $Z(\mathbf{u}_0)$ at the estimation location $\mathbf{u}_0$ as the following linear combination of the data in a string $Z(\mathbf{u}_i), i = 1,\ldots,n,$

$$Z_{FDOK}{}^*(\mathbf{u}_0) = \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{k} \lambda^k_{OK,i}(\mathbf{u}_0) Z^k_i \tag{3.35}$$

where $Z^k = (Z^k_1,\ldots,Z^k_k)$, $k = 1,\ldots,n,$ denotes the vector of k closest data in a string to the estimation location $\mathbf{u}_0$; $\boldsymbol{\lambda}^{k,T}_{OK} = (\lambda^k_{OK,1},\ldots,\lambda^k_{OK,k})^T$, $k = 1,\ldots,n,$ denotes the vector of the ordinary kriging weights calculated from the normal system of equations for the estimation location $\mathbf{u}_0$ based on the k closest data in the string

$$\sum_{i=1}^{k} \lambda^k_{OK,i} Cov(Z^k_j, Z^k_i) + \mu = Cov(Z(\mathbf{u}_0), Z^k_i), \quad j = 1,\cdots,k, \tag{3.36}$$

$$\text{and} \quad \sum_{i=1}^{k} \lambda^k_{OK,i} = 1, \tag{3.37}$$

where μ is a Lagrange multiplier, and, as before, $Cov(Z^k_j, Z^k_i)$ and $Cov(Z(\mathbf{u}_0), Z^k_i)$ denote data-to-data covariance and data-to-estimation point covariance, respectively, $i, j = 1,\cdots,k,\ k = 1,\ldots,n.$

In terms of the finite domain ordinary kriging weights $\boldsymbol{\lambda}^T_{FDOK}(\mathbf{u}_0) = (\lambda_{FDOK,1}(\mathbf{u}_0),\ldots,\lambda_{FDOK,n}(\mathbf{u}_0))^T = (\lambda_{FDOK,1},\ldots,\lambda_{FDOK,n})^T$,

$$\lambda_{FDOK,i} = \frac{1}{n}\sum_{k=1}^{n} \tilde{\lambda}^k_{OK,i}, \tag{3.38}$$

where

$$\tilde{\lambda}^k_{OK,i} = \begin{cases} \lambda^k_{OK,m}, & \text{if } i\text{ - th data is within } k \text{ closest data in the string to the estimation location } \mathbf{u}_0; \\ & \text{index } m \text{ denotes the order of data } Z_i(\mathbf{u}_0) \text{ in the vector } Z^k; \\ 0, & \text{otherwise.} \end{cases}$$

the finite domain ordinary kriging estimator (3.35) can be rewritten as

$$Z_{FDOK}{}^*(\mathbf{u}_0) = \sum_{i=1}^{n} \lambda_{FDOK,i}(\mathbf{u}_0) Z_i(\mathbf{u}_0). \tag{3.39}$$

The estimation variance at the estimation location $\mathbf{u}_0$ for the finite domain ordinary kriging estimator is given by:

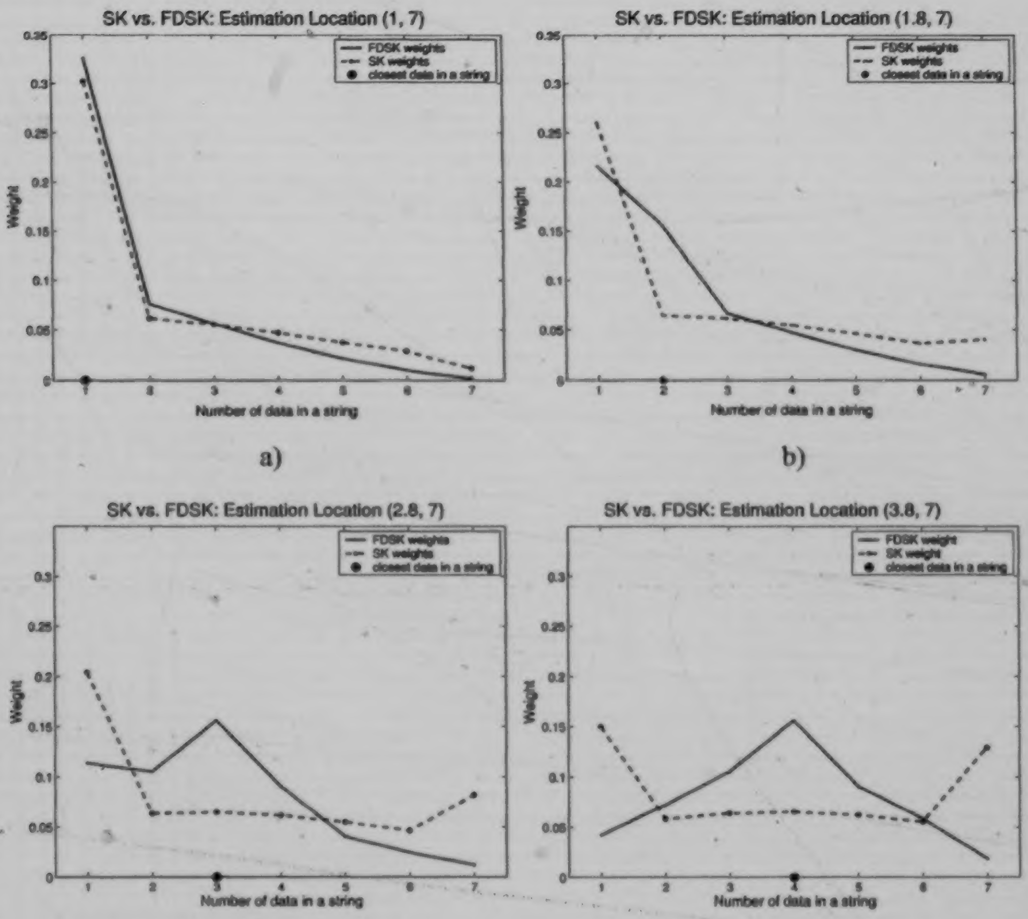
$$\sigma^2_{FDOK}(\mathbf{u}_0) = \sigma^2 - 2\sum_{i=1}^{n} \lambda_{FDOK,i} Cov(Z(\mathbf{u}_i), Z(\mathbf{u}_0)) + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_{FDOK,i}\lambda_{FDOK,j} Cov(Z(\mathbf{u}_i), Z(\mathbf{u}_j)) \tag{3.40}$$

3.4.2. Small Examples

To compare the kriging weights obtained using the traditional kriging (simple and ordinary) with the finite domain kriging approaches several small studies were performed. The weights were calculated for four estimation locations, (1, 7), (1.8, 7), (2.8, 7) and (3.8,7), based on the string of 7 data located at (1,0), (2,0), (3,0), (4,0), (5,0), (6,0) and (7,0), respectively. An isotropic spherical variogram with a sill of one and range of 20 is considered for analysis. The weights are shown in Figure 3.8 for simple kriging and finite domain simple kriging. Figure 3.9 shows ordinary kriging and finite domain ordinary kriging. Note from Figures 3.8-3.9 how finite domain kriging removes the string effect; the artificially higher weights given to the end samples of the string are significantly reduced.

Furthermore, note the smoothness of the finite domain kriging weights compared to distance constrained kriging weights (Figures 3.6-3.7). The distance constrained correction presented in Section 3.3 resulted in very non-smooth weights. The smoothness in the structure of the finite domain kriging weights is a direct result of averaging or taking the expected value of the optimal kriging estimators for different search neighborhoods, no (distance) constrains that may result in sub-optimal solutions are involved.

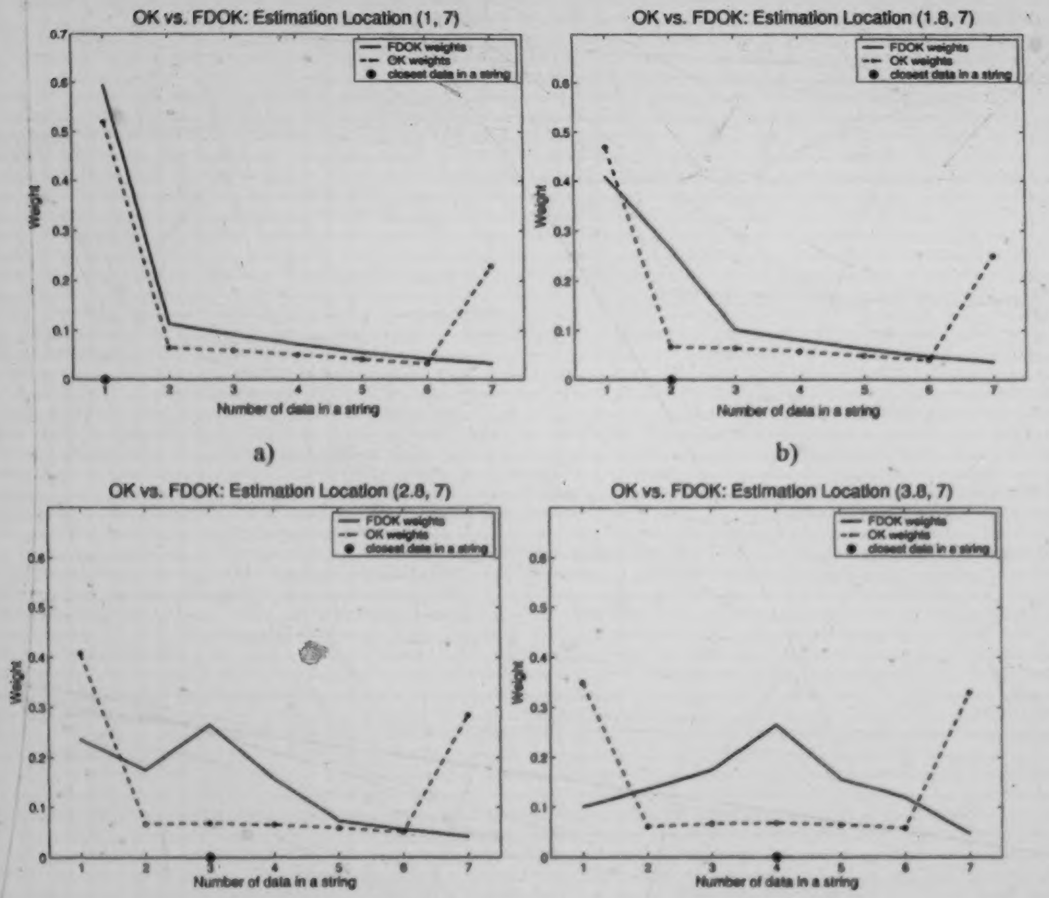
Figure 3.10 shows the results with simple kriging weights with respect to the number of closest data in a string used for estimation. Figure 3.10 also shows the finite domain simple kriging weights resulting from averaging of all simple kriging results for the weights.



a) b)

c) d)

Figure 3.8: Profiles of the finite domain simple kriging (solid line) and simple kriging (dashed line) weights. Results shown for estimation locations: a) (1, 7); b) (1.8, 7); c) (2.8, 7); and d) (3.8, 7) are obtained using a spherical variogram model with a range of 20 based on a string of 7 data located at (1,0), (2,0), (3,0), (4,0), (5,0), (6,0) and (7,0), respectively. The closest data in a string is denoted by a dark circle.



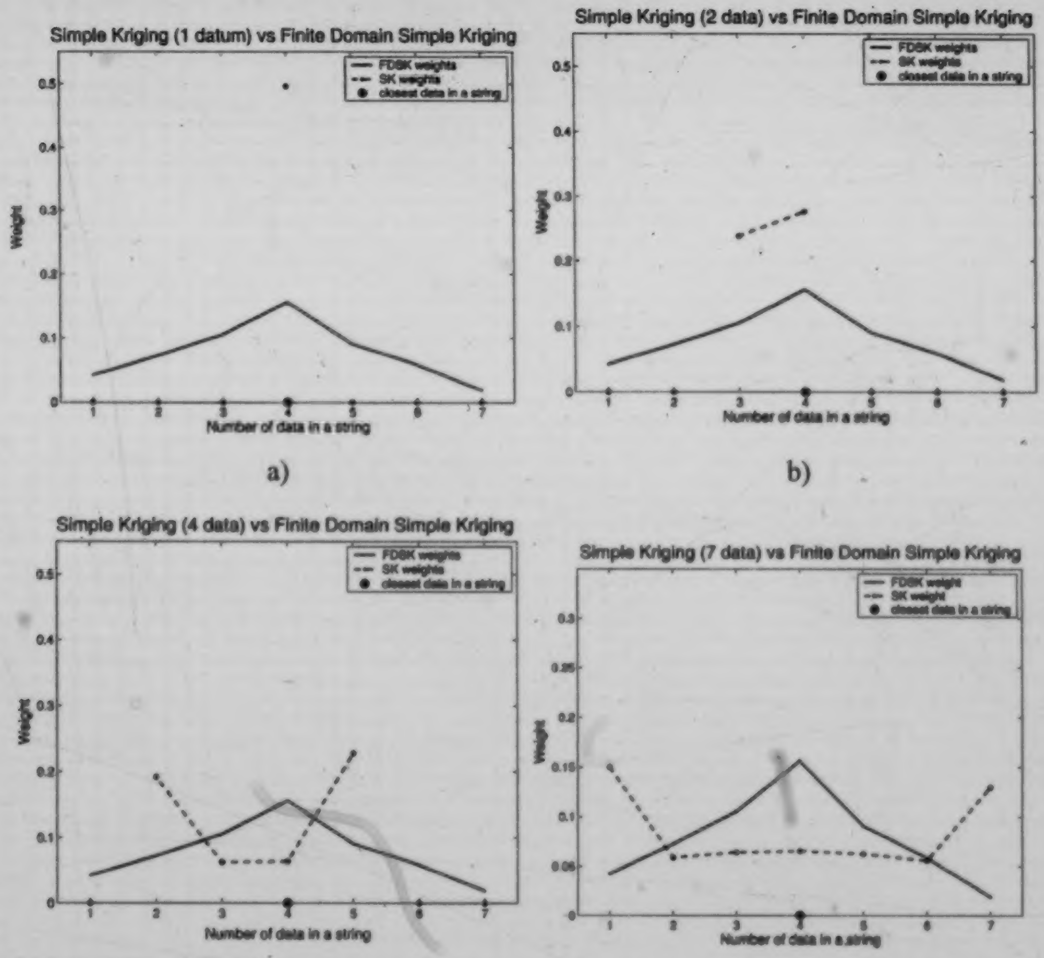
a)

b)

c)

d)

Figure 3.9: Profiles of the finite domain ordinary kriging (solid line) and ordinary kriging (dashed line) weights. Results shown for estimation locations: a) (1, 7); b) (1.8, 7); c) (2.8, 7); and d) (3.8, 7) are obtained using a spherical variogram model with a range of 20 based on a string of 7 data located at (1,0), (2,0), (3,0), (4,0), (5,0), (6,0) and (7,0), respectively. The closest data in a string is denoted by a dark circle.



a) b)

c) d)

Figure 3.10: Change in the structure of the simple kriging weights (dashed lines) with respect to the number of closest data in a string used for estimation of location (3.8,7). Results are shown for: a) 1 data; b) 2 data; c) 4 data; and d) 7 data are obtained using spherical variogram model with the range of correlation 20. Finite domain simple kriging weights calculated based on all 7 data are shown in solid line. String of 7 data is located at (1,0), (2,0), (3,0), (4,0), (5,0), (6,0) and (7,0), respectively. The closest data in a string is denoted by a dark circle.

3.4.3. Properties

Finite domain (simple and ordinary) kriging estimators are linear combinations of several traditional kriging estimators, each of which is unbiased and exact; therefore, we conclude that

1. Finite domain kriging estimators are unbiased.
2. Finite domain kriging estimators are exact.
3. Finite domain kriging estimators provide models for the unsampled value of the variable of interest according to its spatial continuity described by the covariance function. Finite domain kriging takes into account the redundancy of data in the string and closeness of the data in the string to an estimation location.

Moreover, finite domain kriging estimators possess an additional interesting property. Consider estimation at a particular location based on a very long string of data ('infinite'), we can observe that at some point it becomes irrelevant whether you use the whole string of data or just a portion of a particular size. That is, finite domain kriging weights assigned to the 'infinite' string of data will be virtually the same as assigned to a long substring and zero weights for the rest of the data in the string. Mathematically we can write this property as follows:

There exists number l of closest data in the 'infinite' string to the estimation location such that the following inequality holds

$$\left[\sum_{i=k}^{k+l}(\lambda_{FDK,i}-\widetilde{\lambda}_{FDK,i})^2+\sum_{i<k \text{ or } i>k+1}\lambda_{FDK,i}{}^2\right]<\varepsilon, \qquad (3.41)$$

where $\boldsymbol{\lambda}_{FDK}^T=(\lambda_{FDK,1},\lambda_{FDK,2},\ldots,\lambda_{FDK,n})^T$ *is the n by 1 vector of the finite domain kriging weights assigned to the 'infinite' string,* $\widetilde{\boldsymbol{\lambda}}_{FDK}^T=(\widetilde{\lambda}_{FDK,k},\widetilde{\lambda}_{FDK,k+1},\ldots,\widetilde{\lambda}_{FDK,k+l})^T$ *is the l by 1 vector of finite domain kriging weights assigned to the l closest data in the string to the estimation location; and* ε *is a very small number, say* $\varepsilon=0.0001$.

Further we will refer to property Equation (3.41) of finite domain kriging as the convergence property. To illustrate the convergence property of finite domain kriging, the following small case study was conducted. A string of 3000 data located at (1,0), (2,0), …, (3000,0), respectively, was considered for estimation of location (100,7) using finite domain kriging techniques based on a spherical variogram model with a range of correlation 500. Finite domain kriging was performed using 25, 100, 250, 500, 1000, 1500 and 3000 closest samples in the string. Figure 3.11 shows the resulting change in the structure of the finite domain simple kriging weights with respect to the number of closest data. Note that there is virtually no difference in finite domain simple kriging weights when performing estimation based on 1500 (half string) or more data or full 'infinite' string. Specifically, the difference in the left hand side of inequality (3.41) for finite domain simple kriging weights calculated based on the full string of 3000 data and 1500 data is less than 5.1169e-006. Moreover, the result of finite domain simple kriging estimation based on 500 or more data and the whole string of 3000 data will be also very similar, see Figure 3.11.

The difference in the left hand side of inequality (3.41) for all cases of finite domain simple kriging estimation are given in Table 3.1.

Table 3.1: The value in the left hand side of inequality (3.41) as a function of the number of data used in the finite domain simple kriging estimation.

25 data	100 data	250 data	500 data	1000 data	1500 data
0.0061	0.0025	0.0023	0.0008	0.0007	0.0000

Figure 3.12 shows the change in the structure of the finite domain ordinary kriging weights with respect to the number of closest data. There is virtually no difference in finite domain ordinary kriging weights when performing estimation based on 1500 (half string) or the full 'infinite' string. Specifically, the difference in the left hand side of inequality (3.41) for finite domain ordinary kriging weights calculated based on the full string of 3000 data and 1500 data is less than 5.2874e-006.

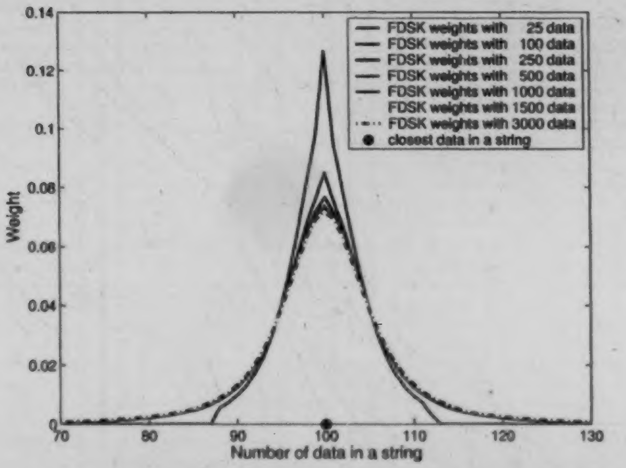


Figure 3.11: Change in the structure of the finite domain simple kriging weights with respect to the number of closest data in a string used for estimation of location (100,7). Results are obtained using spherical variogram model with the range of correlation 500. The string of 3000 data is located at (1,0), (2,0), …, (3000,0), respectively.

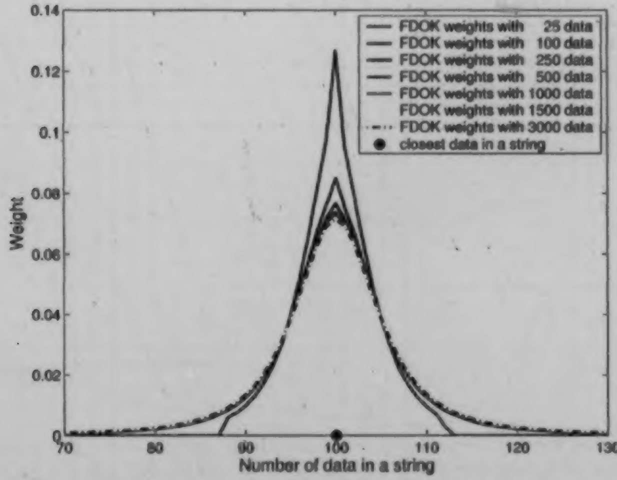


Figure 3.12: Change in the structure of the finite domain ordinary kriging weights with respect to the number of closest data in a string used for estimation of location (100,7). Results are obtained using spherical variogram model with the range of correlation 500. The string of 3000 data is located at (1,0), (2,0), …, (3000,0), respectively.

Again note that the result of the finite domain ordinary kriging estimation based on 500 or more data and the whole string of 3000 data will be very similar, see Figure 3.12. The difference in the left hand side of inequality (3.41) for all cases of finite domain ordinary kriging estimation are given in the Table 3.2.

Table 3.2: The value in the left hand side of inequality (3.41) as a function of the number of data used in the finite domain ordinary kriging estimation.

25 data	100 data	250 data	500 data	1000 data	1500 data
0.0062	0.0025	0.0023	0.0008	0.0007	0.0000

Note that the results of finite domain simple kriging and finite domain ordinary kriging shown in Figures 3.11-3.12 are very similar; this is because simple kriging and ordinary kriging have very similar weights due to the closeness of the estimation location to the string in terms of the variogram range of continuity.

3.4.4. Finite Domain Kriging: Generalization to the Case of Multiple Strings

Let us consider the situation of multiple strings containing possibly different numbers of data. Recall that the finite domain kriging estimator in the single string case is an average of several traditional kriging (simple or ordinary) estimators obtained using different neighborhood search strategies. When there are multiple strings they can be considered together at the same time or separately. This leads to two types of finite domain kriging.

Finite domain kriging I performs the traditional kriging as many times as there are conditioning data, that is, n. Each time k, $k=1,\dots,n$, kriging is performed based on the k closest data without considering if they are from different strings or the same string. Thus, basically, the procedure of the finite domain kriging I is the same as described for the single string case.

Finite domain kriging II is slightly more complicated. Assume first that each string l, $l=1,\dots,L$, contains at least n data. Then, in order to obtain finite domain kriging II estimate, the traditional kriging is performed n times. Each time k, $k = 1,\ldots;n$, kriging is performed based on a set of k closest data from each string. If the strings contain a different number of data, say string j contains only m ($m < n$) data, then the procedure is almost the same. Except that in order to obtain final finite domain kriging II, each time k, $k = 1,\ldots,m$, kriging is performed based on a set of k closest data from each string; while each time k, $k = m+1,\ldots,n$, kriging is performed based on a set of k closest data from all strings expect for string j (from string j only m data are selected). This procedure could be easily extended to the case with a different number of data in each string.

3.5. Comparison of Distance Constrained Kriging and Finite Domain Kriging

To assess the performance of the distance constrained kriging and finite domain kriging in correcting the string effect, a case study of the real data from a petroleum reservoir is conducted. In total, there are 180 vertical wells with information on bitumen in the study area of 1800 by 2500 by 130 meters. The locations of the wells in the *XY* plane are shown in Figure 3.13. Figure 3.13 also shows the histogram of the data.

A subset of 90 wells is chosen from 180 wells to evaluate the difference in the estimation results produced by ordinary kriging, distance constrained kriging and finite domain ordinary kriging, see Figure 3.13. These wells will be used in estimation of the study domain as well as in checking the kriging estimates at the validation well (the other 90 wells) locations. In accordance with standard practice all analysis is conducted in normal score units.

Figure 3.14 shows the experimental variogram and a theoretical fit in the horizontal directions of major and minor continuity and the vertical direction. Variograms are calculated based on normal score transformed data from the 90 estimation wells.

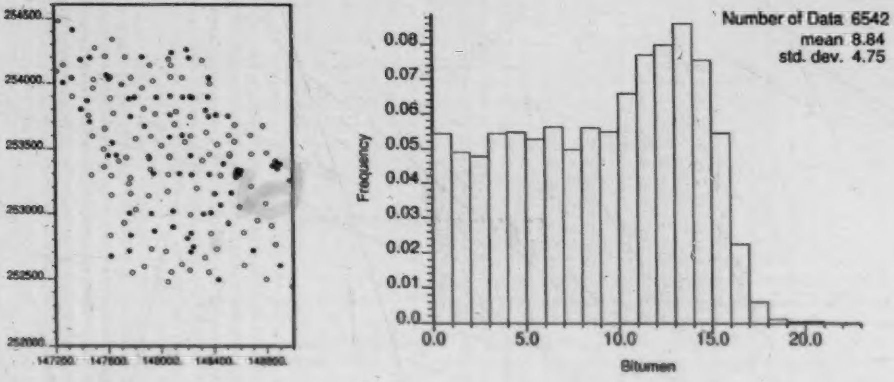
Figure 3.15 shows the difference in the estimates obtained in DCOK and OK and the difference in the estimates obtained in FDOK and OK for the slice at 265 m with respect to the vertical direction of the 3D model for the normal score transformed bitumen. Note that the neigbourhood search for each kriging procedure was specified based on the variogram model of the data (that is, maximum seach radii were set to the largest variogram ranges in the respective directions); min and max data for kriging was set to 10 and 20, respectively. When comparing the maps shown in Figure 3.15 we can clearly note that the locations of areas of low and high values are the same. These areas indicate the areas affected by the string effect of kriging. Moreover, note that Finite Domain Ordinary Kriging seems to result in a stronger correction of the string effect, see Figure 3.15.

The correction of the string effect provided by the two finite domain approaches presented in this paper is a local correction. The global mean squared error (MSE) obtained in Ordinary Kriging, Finite Domain Ordinary Kriging and Distance Constrained Kriging are very close. DCOK and FDOK perform better than OK in terms of MSE by 4.56% and 1.17%, respectively. However, when we compare local changes where the estimates are significantly different, the situation is dramatically different.

Different estimates are identified when the absolute difference between absolute residuals of ordinary kriging and FDOK or DCOK exceeds some cut-off. The value of this absolute difference will be denoted by Δ. Table 3.3 shows results of DCOK vs. OK and FDOK vs. OK with respect to the percentage of estimates affected by the strong string effect; percentage improvement in MSE for $\Delta > 0.1$, and $\Delta > 0.2$.

Table 3.3: Performance of OK vs. FDOK and DCOK in jackknife.

	$\Delta > 0.1$		$\Delta > 0.2$	
Methods	% of Estimates	% Improvement in the MSE	% of Estimates	% Improvement in the MSE
DCOK vs. OK	33.48%	9.99%	13.97%	16.13%
FDOK vs. OK	20.02%	3.31%	5.03%	6.08%



a) b)

Figure 3.13: Locations of 180 wells with data (a) together with bitumen distribution (b). Estimation wells are shown in dark circles.

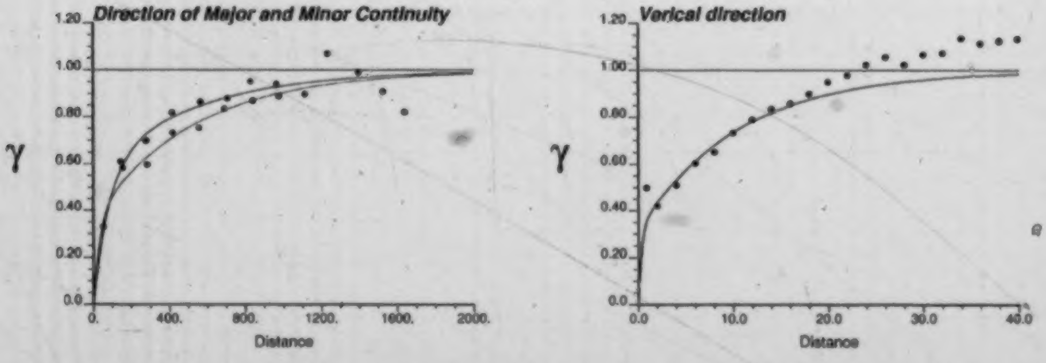
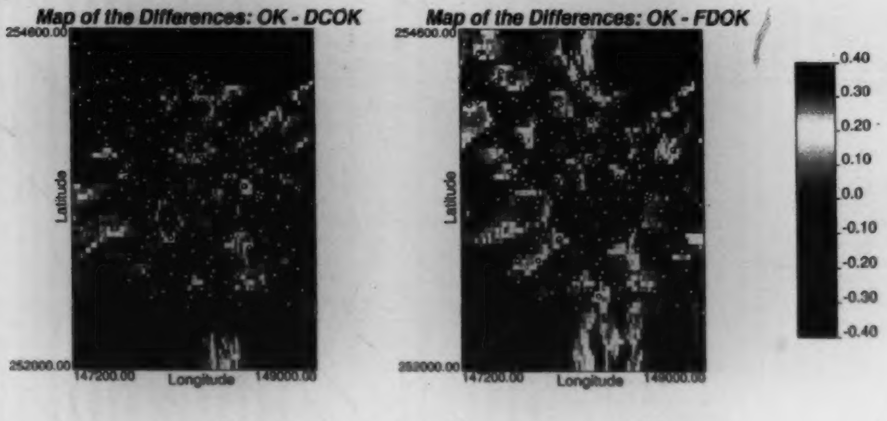


Figure 3.14: Experimental variogram and its theoretical fit in the three directions of major continuity for the normal score transformed bitumen from 90 estimation wells.


Map of the Differences: OK - DCOK
Map of the Differences: OK - FDOK
254600.00
252000.00
Latitude
147200.00
Longitude
149000.00
0.40
0.30
0.20
0.10
0.0
-0.10
-0.20
-0.30
-0.40

a) b)

Figure 3.15: Maps of the differences between OK and DCOK (a) and OK and FDOK (b) for the slice at 265 m of the 3D model for the normal score transformed bitumen.

Note that with an increase in Δ, the outperformance of the two finite domain kriging approaches over ordinary kriging increases, while the percentage of the data affected by the strong string effect decreases. Note also that FDOK results in stronger correction; however, DCOK performs better in jackknife validation. The situation could be different in different areas or with different data.

3.6. Discussion

Two new approaches for estimation in a finite domain using strings of data are presented: distance constrained kriging and finite domain kriging. Both methods result in linear unbiased estimators. Distance constrained kriging is obtained by minimizing the estimation variance subject to distance constraints. Finite domain kriging is obtained as an average (or expected) value of the optimal kriging estimators for different search neighborhoods. Both new methods are applicable when the data at the end of strings of data are somehow anomalous, for example, thin deposits with vertical trends.

Finite domain kriging and distance constrained kriging were evaluated with a case study. Both approaches were shown to provide improved estimates. In this particular case study, finite domain kriging was shown to result in stronger correction; however, distance constrained kriging performed better in jackknife validation. The situation could be different in different areas or with different data sets.

CHAPTER 4

Uncertainty as the Overlap of Distributions

This chapter investigates a new approach for combining local distributions of uncertainty obtained from different data sources or interpolation techniques. The proposed approach leads to narrower but fair local uncertainty.

Section 4.1 reviews results of different comparative studies aimed at finding the best technique and explains the need for a combined distribution. **Section 4.2** presents a new approach, referred to as the overlap uncertainty, for combining alternate conditional distributions of uncertainty. Application of the overlap uncertainty approach is shown in **Section 4.3** by combining inverse distance interpolation results with simple kriging. Further analysis of the overlap uncertainty approach is conducted in **Section 4.4** based on a small case study. A brief discussion about the results and the method is presented in **Section 4.5**.

4.1. Introduction

An important task in modern geostatistics is the assessment and quantification of resource uncertainty. This uncertainty is valuable support information for many management decisions. Uncertainty at specific locations and uncertainty in the global resource are of interest. There are many different methods/interpolation techniques to build models of uncertainty including kriging, cokriging, inverse distance, etc. Each method leads to different results.

The variety of available interpolation techniques has led to a question about which is the best or the most appropriate method (Voltz and Webster, 1990; Laslett, 1994; Hosseini et al., 1994; Wollenhaupt et al., 1994; Gotway et al., 1996). The search for the optimum interpolation technique has resulted in a large number of comparative studies based on realistic or geologically sound visual appearance; resampling with cross validation or the jackknife (Isaaks and Srivastava, 1989); robustness; or a measure of response variables derived from the interpolated property. Robustness refers to a solution that is stable with respect to variations in the input parameters (Falivene et. al., 2007). Most of these studies have resulted in diverse conclusions. In some studies, geostatistical kriging-based methods performed best (Creutin and Obled, 1982; Tabios and Salas, 1985; Rouhani, 1986; Grimm and Lynch, 1991; Laslett and McBratney, 1990; Weber and Englund, 1994; Laslett, 1994; Phillips et al., 1997; Borga and Vizzaccaro, 1997; Zimmerman et al., 1999; Goovaerts, 2000; Teegavarapu and Chandramouli, 2005), while in others, inverse distance weighting or spline methods were as good or better (Weber and Englund, 1992; Gallichand and Marcotte, 1993; Boman et al., 1995; Brus et al, 1996; Declercq, 1996; Borga and Vizzaccaro, 1997; Dirks et al., 1998; Moyeed and Papritz, 2002). Moreover, even among kriging estimators it is impossible to select a unique best or optimal estimator (Zimmerman, et. al., 1999; Moyeed and Papritz, 2002).

Martinez-Cob (1996), Caruso and Quarta (1998) and Nalder and Wein (1998) established that the differences in the performance among various spatial interpolation methods are largely influenced by features of the spatial variable (that is, skewness and kurtosis of the data; the coefficient of variation; and whether the data contains extreme observations (Schloedar et. al., 2001), spatial configuration of the data and assumptions required by the method, rather than the method of spatial interpolation itself. The above-mentioned effects can not be removed from the results; and, thus, different characteristics of the spatial variable may lead to different performance and reliability measures (Mardikis et. al., 2005).

Despite of all the comparative studies and the search for the optimal interpolation strategy, little or no attention has been put towards combining the good features of the approaches. Each technique is good in different senses, that is, simplicity, robustness, reliability, flexibility, geological realism, statistical accuracy, etc.; while none of these techniques, can be regarded as the best one in general (Rana and Katerji, 2000). Therefore, a method for combining spatial predictions and associated models is of interest. One method for combining local uncertainties is developed below. This method is referred to as an overlap uncertainty estimator. The new estimator is obtained as an overlap of alternate conditional distributions.

4.2. Uncertainty as the Overlap of Distributions

Let us consider *n* data on a spatial random variable *Z* at locations $\mathbf{u}_i$, $i = 1,\ldots,n$, in the domain of interest *A*. Consider the problem of estimating the value of the variable *Z* at an unsampled location $\mathbf{u}_0$ and quantification of the local uncertainty at this location. Consider two different techniques available to obtain local uncertainty at the location $\mathbf{u}_0$. Figure 4.1 shows a schematic representation of the results for the local uncertainty at location $\mathbf{u}_0$. Note that the results are different for each estimation technique. To reconcile the local uncertainty distributions, we propose to model the local uncertainty of *Z* at an unsampled location $\mathbf{u}_0$ as the minimum or overlap of the local uncertainties obtained by two different estimation techniques (scaled to 1). The uncertainty model obtained as an overlap of the two modeled local uncertainties is often narrow and appears as a reasonable result. An assumption is that the probability is the highest that the true value at the unsampled location is in the interval common to both estimators. This estimator will be referred to as an overlap estimator. Figure 4.1 shows schematic representation of the overlap of the two local uncertainty models at location $\mathbf{u}_0$.

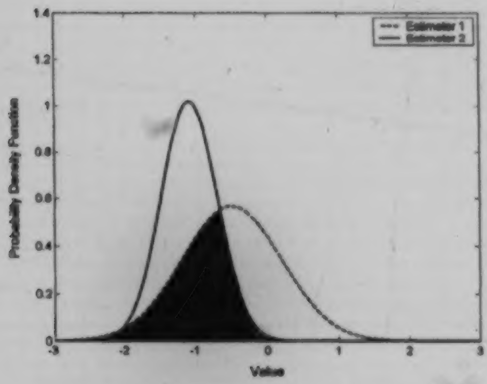


Figure 4.1: Schematic representation of the example results for the two local uncertainty at location $\mathbf{u}_0$. Solid and dashed lines represent local uncertainty obtained by two different estimation approaches; dark area is an overlap.

This approach can easily be extended to the case of more than two estimators. In particular, simple kriging, ordinary kriging, inverse distance, ckriging, Bayesian udating, etc., all can be considered. The overlap estimator can take on any arbitrary shape, see schematic examples in Figure 4.2.

4.3. Example with Inverse Distance and Simple Kriging

To illustrate the idea of the overlap uncertainty estimator, we limit ourselves to a standard normal spatial random variable and two techniques for its estimation, that is, inverse distance and simple kriging. A short description of the simple kriging can be found in Chapter 2 of this thesis, a short description of the inverse distance interpolation technique follows.

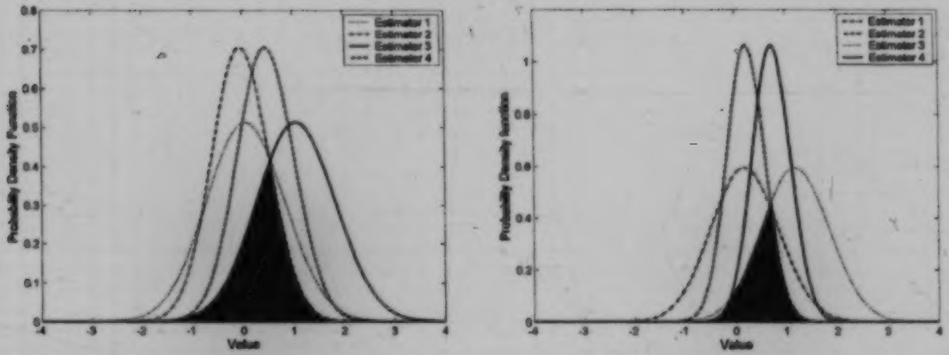


Figure 4.2: Schematic representation of the two example results for the four local uncertainty at location $\mathbf{u}_0$. Solid dashed, doted and dash-dot lines represent local uncertainty obtained by four different estimation approaches; dark area is an overlap.

4.3.1 Inverse Distance Interpolator

An inverse distance (ID) weighted estimate of the variable of interest Z at an unsampled location $\mathbf{u}_0$ is a spatially weighted average of the sample values within a search neighborhood (Shepard, 1968; Franke, 1982; Diodato and Ceccarelli, 2005). It is calculated as

$$Z_{ID}(\mathbf{u}_0) = \sum_{i=1}^{n} \beta_{ID,i} Z(\mathbf{u}_i), \tag{4.1}$$

where $\beta_{ID,i}, i = 1,\ldots,n,$ are the ID weights assigned to each sample:

$$\beta_{ID,i} = \frac{\left(\dfrac{1}{d_i^p}\right)}{\sum_{i=1}^{n}\left(\dfrac{1}{d_i^p}\right)}, \quad (i = 1,\ldots,n), \tag{4.2}$$

where $d_i, i = 1,\ldots,n,$ are the Euclidian distances between estimation location and sample points, and exponent p is the power or distance exponent value. Note that the sum of the inverse distance is one, that is,

$$\sum_{i=1}^{n} \beta_i = 1. \tag{4.3}$$

The most common value applied for the power p is 2; then estimator in (4.1)-(4.3) is called inverse squared distance (ISD) interpolator. However, any value for p can be chosen. As p increases, the interpolated value by inverse distance is assigned the value of the nearest sample point, that is, inverse distance estimate becomes the same as estimate produced by polygonal method. (Diadato and Ceccarelli, 2005).

The mean and variance of the inverse distance estimator $Z_{ID}(\mathbf{u}_0)$ at estimation location $\mathbf{u}_0$ under an assumption of stationarity and multivariate Gaussianity can be derived as follows

$$E(Z_{ID}(\mathbf{u}_0)) = E\left(\sum_{i=1}^{n}\beta_{ID,i}Z(\mathbf{u}_i)\right) = \sum_{i=1}^{n}\beta_{ID,i}E(Z(\mathbf{u}_i)) = m\sum_{i=1}^{n}\beta_{ID,i} = m; \qquad (4.4)$$

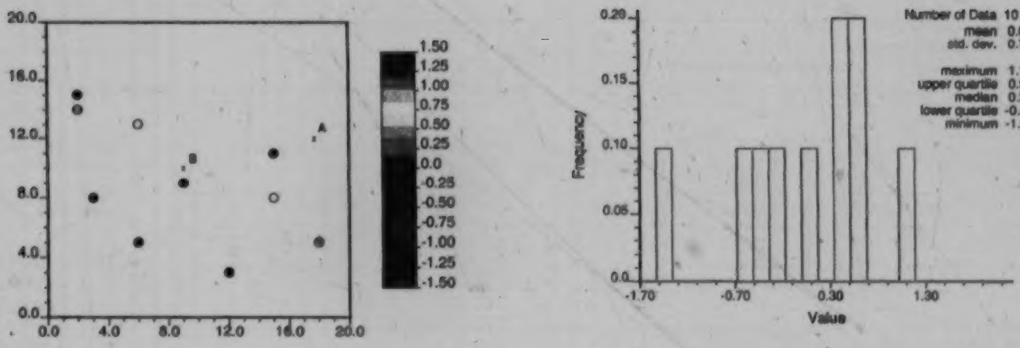
$$Var(Z_{ID}(\mathbf{u}_0)) = Var\left(\sum_{i=1}^{n}\beta_{ID,i}Z(\mathbf{u}_i)\right) = \sum_{i=1}^{n}\sum_{j=1}^{n}\beta_{ID,i}\beta_{ID,j}Cov(Z(\mathbf{u}_i),Z(\mathbf{u}_j)); \qquad (4.5)$$

where m is stationary domain mean and $Cov(Z(\mathbf{u}_i),Z(\mathbf{u}_j)) = C(\mathbf{u}_i,\mathbf{u}_j),\ \ i,j = 1,\cdots,n,$ denotes data-to-data covariance.

4.3.2. Uncertainty Overlap

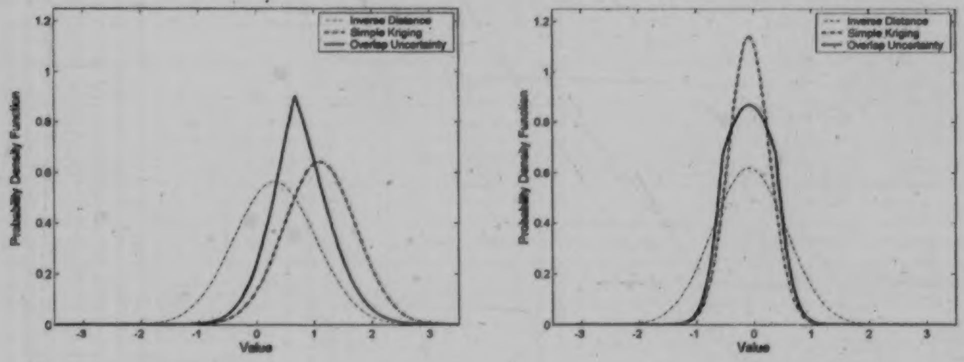
Let us consider 10 simulated data in a domain 20 by 20 units shown in Figure 4.3. The data were generated using a zero nugget spherical variogram model with range of correlation equal to the size of the domain. The data distribution is also shown in Figure 4.3. Note that this 2-D example was chosen small enough to easily visualize the results yet large enough to show realistic variations in the results. The conclusions drawn from this example are considered reasonably general.

Let us now determine the local uncertainty at the two estimation locations $A(18,12)$ and $B(9,10)$ using inverse distance interpolation and simple kriging, then subsequently calculate the overlap uncertainty estimator as the overlap of the two alternate conditional distributions. Figure 4.4 shows results obtained for the two estimation locations.



a) b)

Figure 4.3: Location map of 10 data (a) and their distribution (b). Points *A* and *B* represent two estimation locations of interest.



a) b)

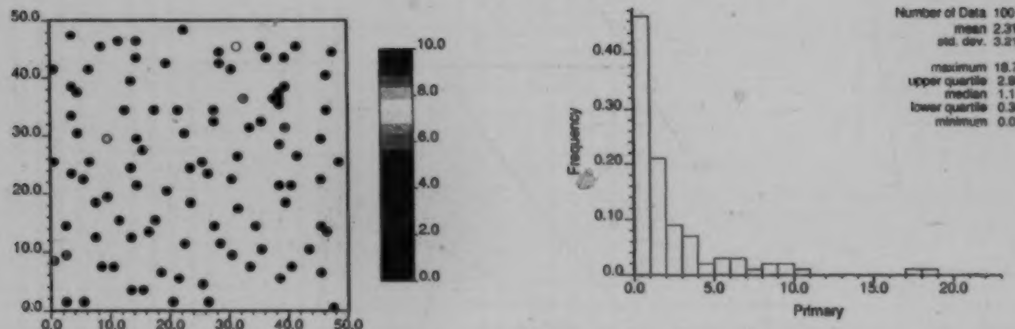
Figure 4.4: (Scaled) overlap uncertainty estimator together with inverse distance and simple kriging local uncertainty models for estimation location *A* (a); and *B* (b). Dash-dot, dashed and solid lines represent local uncertainty obtained by simple kriging, inverse distance and overlap uncertainty approaches.

Looking at Figure 4.4, note that the local uncertainty predicted by the inverse distance interpolation and simple kriging for estimation location $A(18,12)$ are quite different. As a result, when we combine both techniques to obtain the overlap uncertainty estimator for unsampled location $A(18,12)$, we observe that the local uncertainty by this approach is narrower than predicted by either of the two estimation approaches. To quantify the change in the local uncertainty at the estimation location $A(18,12)$ more precisely, we will use the width of the P_{10} to P_{90} interval. For the inverse distance estimator this interval is (-0.59, 1.23); for the simple kriging estimator it is (0.31,1.89) and for the overlap uncertainty estimator it is (0.11,1.45). The difference is quite significant.

On the other hand, the local uncertainty predicted by the inverse distance interpolation and simple kriging for estimation location $B(9,10)$ are similar (see Figure 4.4). As we combine both techniques to obtain the overlap uncertainty for unsampled location $B(9,10)$, we observe that the local uncertainty by this approach is similar to local uncertainty predicted by simple kriging. Specifically, for the inverse distance estimator (P_{10}, P_{90}) probability interval for the estimation location $B(9,10)$ is (-0.90,0.75); for simple kriging estimator it is (-0.59,0.37) and for the overlap uncertainty estimator it is (-0.53,0.44).

4.4. Case Study

To illustrate the performance of the overlap uncertainty estimator the well known GSLIB (Deutsch and Journel, 1998) data set 'cluster.dat' is selected. The data consists of 100 data that are sampled on a random stratified grid and 40 data that are clustered in high valued areas. We discard the clustered data. The 2-D area of interest is 50 by 50 distance units. The distribution of data is approximately lognormal with a mean of 2.5 and a standard deviation of 5.0. The spatial continuity of the data in the normal space is described by an isotropic spherical variogram model with a range of correlation of 12 and nugget effect of 0.3. Figure 4.5 shows the location map of the 100 data and their distribution.



a)

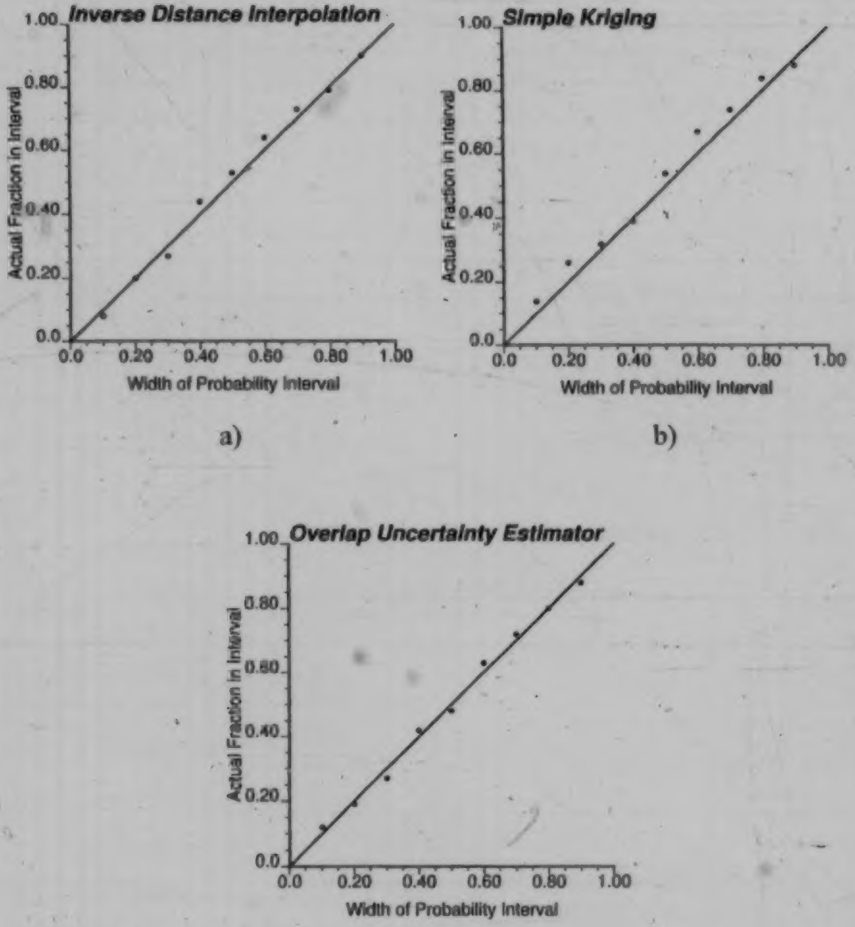
b)

Figure 4.5: Location map of 100 data from file 'cluster.dat' (a) and their distribution (b).

The aim of our analysis is to establish local conditional distributions for all 100 data points in cross validation mode and check the results. Specifically, the fairness and accuracy of the local uncertainties will be checked as well as the errors in estimation. For simplicity, all analysis is conducted in normal space.

Figure 4.6 shows accuracy plots (Deutsch, 1996) obtained for inverse distance, simple kriging and overlap uncertainty estimator predictions in cross validation of 100 normal score transformed data of Figure 4.5. From Figure 4.6 we can clearly see that overlap uncertainty estimator is both accurate and precise, thus it is a fair estimator of uncertainty. Both simple kriging and inverse distance are accurate, but less precise estimators of uncertainty.

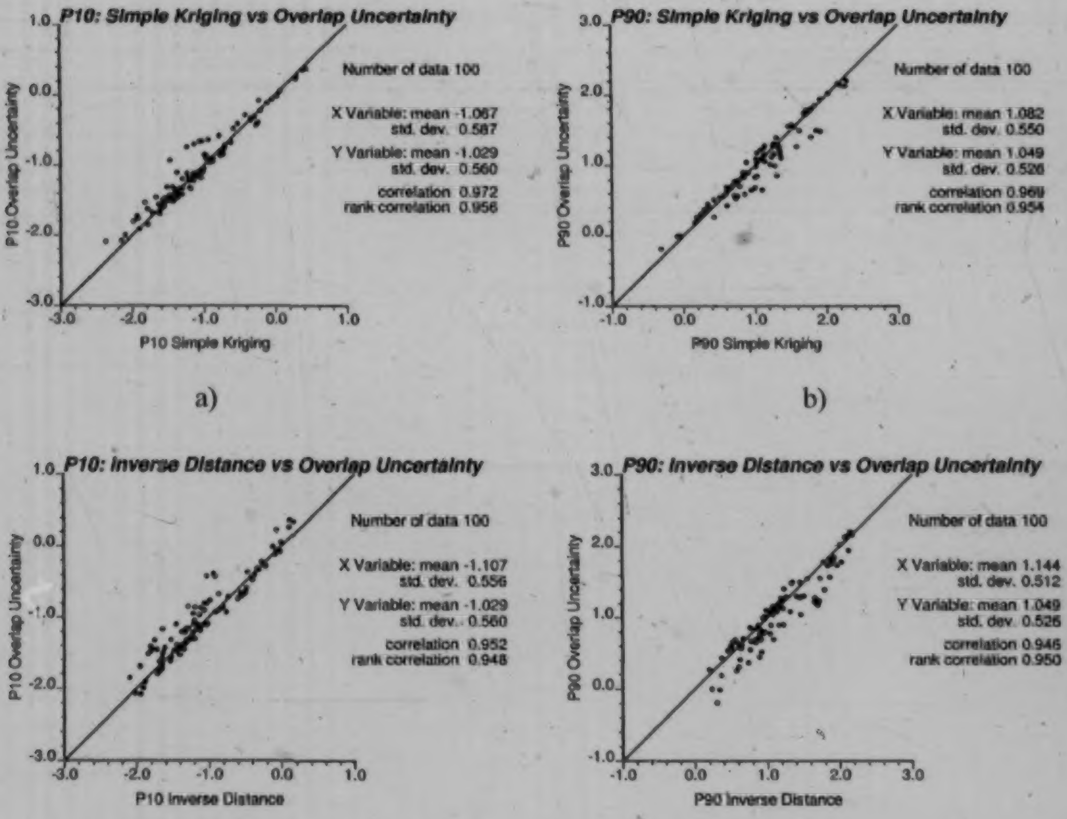
Figure 4.7 shows the crossplots between P_{10} and P_{90} of simple kriging and P_{10} and P_{90}, respectively, of overlap uncertainty estimator and crossplots between P_{10} and P_{90} of inverse distance and P_{10} and P_{90}, respectivey, of overlap uncertainty for 100 normal score transformed data. Figure 4.8 shows the (P_{10}, P_{90}) local uncertainty intervals for the first 10 data. Note that taking uncertainty as the overlap of the local conditional distributions from simple kriging and inverse distance (for example) can lead to significant reduction of the probability intervals (see Figure 4.7). Nevertheless, the uncertainty is not overly constrained; it is still fair.



a) b)

c)

Figure 4.6: Accuracy plots for inverse distance (a), simple kriging (b), and overlap uncertainty estimator (c). Results are obtained in crossvalidation of 100 normal score transformed primary data from file 'cluster.dat'.



a)

b)

c)

d)

Figure 4.7: Crossplots between: P_{10} (a) and P_{90} (b) of the simple kriging and overlap uncertainty estimator and crossplots between P_{10} (c) and P_{90} (d) of the inverse distance and overlap uncertainty estimator for 100 normal score transformed primary data from file 'cluster.dat'.

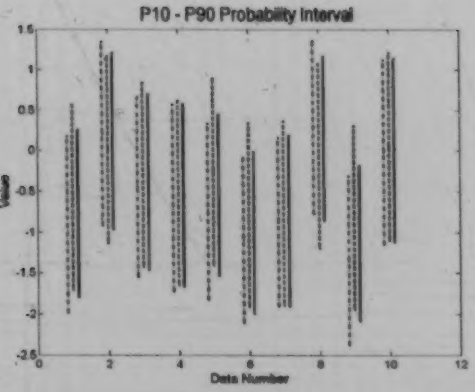
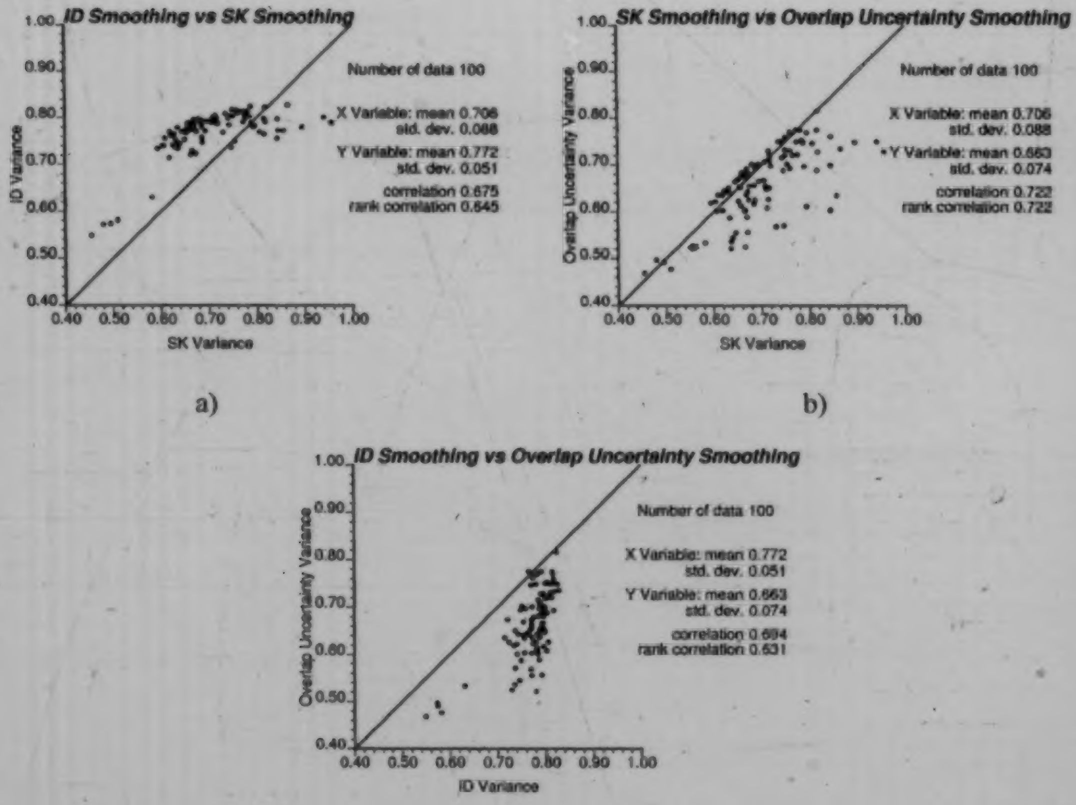


Figure 4.8: (P_{10}, P_{90}) probability intervals obtained for the first 10 data in the 'cluster.dat' data set based on the simple kriging (dash-dot lines), inverse distance interpolation (dashed lines) and overlap uncertainty estimator (solid lines). Medians (P_{50}) for each of the three considered approaches are shown by dots.

To assess how much narrower are local conditional distributions predicted by overlap uncertainty versus simple kriging and inverse distance, crossplots between the variance of the local conditional distributions (smoothing effect) of the simple kriging, inverse distance and overlap uncertainty estimators for the 100 data of the file 'cluster.dat' are prepared. Figure 4.9 shows the results. We observe that the average variance of the local conditional distributions obtained by the simple kriging is 0.706, by the inverse distance is 0.772 and by the overlap of the local conditional distributions is 0.663. Thus, we see that the average variance of the local conditional distributions of the overlap uncertainty estimator is, on average, significantly smaller than that of simple kriging (more than 6% smaller) and that of inverse distance (more than 16% smaller).

Figure 4.10 shows the true values versus estimates from cross validation for inverse distance, simple kriging and overlap uncertainty estimator for 100 normal score transformed 'cluster.dat' data. Looking at Figure 4.10 we can see that all three estimators are virtually unbiased; the correlation between true values and estimates are the highest for the simple kriging approach, and only slightly lower for the overlap uncertainty method.

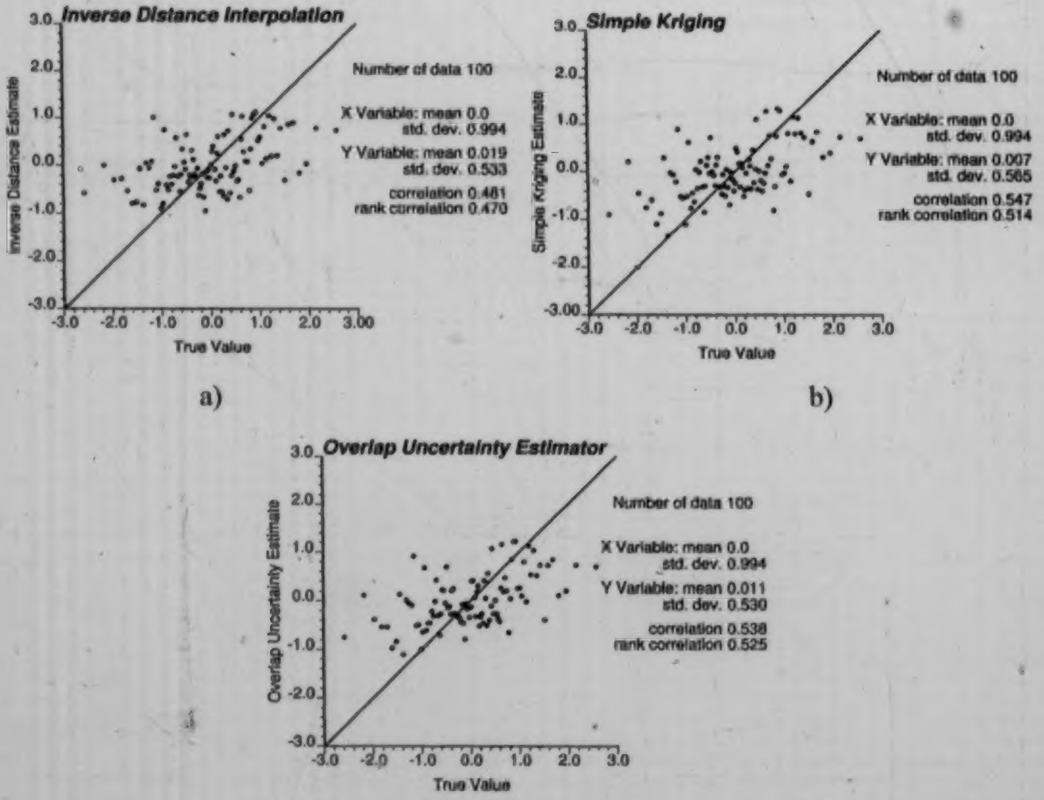


a)

b)

c)

Figure 4.9: Crossplots between the variance of the local conditional distributions (smoothing effect) of the: inverse distance and simple kriging (top left); inverse distance and overlap uncertainty estimator (top right) and simple kriging and overlap uncertainty estimator (bottom) obtained for the 100 data of the file 'cluster.dat'.



a)

b)

c)

Figure 4.10: Cross validation results for the inverse distance interpolation (a); simple kriging (b); and overlap uncertainty estimator (c) obtained for 100 normal score transformed primary data from file 'cluster.dat'.

4.5. Discussion

A flexible approach to combine alternate local conditional distributions to create an overlap uncertainty estimator was proposed. This approach was illustrated using overlap between simple kriging and inverse distance methods in several case studies. It was shown that overlap uncertainty estimator can result in significantly narrower intervals for the local uncertainty; while it does not overly constrain uncertainty. Uncertainty obtained by overlap uncertainty approach will be fair provided the local distributions obtained by the approaches for combine are accurate. Otherwise, accuracy of the local distributions obtained using overlap uncertainty approach should not be expected. However, narrower (on average) local distributions with the approximate accuracy level as the approaches used in overlaping should usually be obtained.

Note also that the overlap uncertainty approach is only aimed at combing alternate local conditional distributions of continuous variables with an infinite support. In should not be applied to discrete variables or continuous variables with a finite support.

CHAPTER 5

Intrinsic Collocated Cokriging

This chapter investigates the sources of variance inflation in collocated cokriging. A new approach for cosimulation of dependent random functions without inferring and modeling of a full cross-covariance matrix is proposed. The proposed method is shown to remove systematic bias in histogram reproduction from the conventional Markov model for collocated cokriging.

Section 5.1 reviews developments in collocated cokriging. **Section 5.2** investigates the theoretical justification for selecting only one auxiliary sample (collocated) for cokriging in the case of an intrinsic correlation model. **Section 5.3** explains the reasons for variance inflation in collocated cokriging. Then an intrinsic collocated cokriging to solve the variance inflation problem is presented. Example applications are given in **Section 5.4. Section 5.5** compares intrinsic collocated cokriging, collocated cokriging and cokriging with a linear model of coregionalization. Comparison is made in terms of the difference in the cokriging weight profiles and results of estimation. **Section 5.6** presents an approach for improved implementation of collocated cokriging in the case of multiple secondary data. The theoretical validity of this approach is proven. **Section 5.7** extends the proposed approach to intrinsic collocated cokriging. Finally, a brief discussion about the results and the methods is presented in **Section 5.8**.

5.1. Introduction

Sequential Gaussian simulation with collocated simple cokriging is a popular method for modeling a primary variable based on extensively sampled secondary information (Xu, *et al.*, 1992; Almeida and Journel, 1994; Goovaerts, 1997). It is widely used because it is simple; the correlation coefficient between the primary variable being modeled and secondary data is the sole additional statistic required to integrate the secondary data.

A number of researchers have explored variants of the original proposal of collocated cokriging (Xu *et al.*, 1992; Almeida and Journel, 1994). Journel explored an alternative Markov assumption (Journel, 1999; Shmaryan and Journel, 1999). The so-called Markov model II or MMII places emphasis on the secondary variable rather than the primary. The MMII model has not gained wide usage because of more complex parameterization and implementation. A multi-collocated cokriging was proposed by Haas *et al.* (1998) (Rivoirard, 2001). The paper of Rivoirard (2001) investigates the two options available for collocated cokriging, that is, collocated simple cokriging and multi-collocated cokriging. The models in which the simplification resulting from the collocated forms does not result in any loss of information are presented by Rivoirard (2001).

Collocated cokriging has a longstanding problem with variance inflation that leads to a systematic bias in the mean and variance of the simulated realizations. An ad hoc method of variance correction has been proposed for dealing with this problem; however, the correction is case dependent and requires manual tuning (Deutsch and Journel, 1998). This tuning is often not performed leading to biased resource estimates.

The aim of this chapter is to explain the theoretical basis for the variance inflation and present an alternative approach to cosimulation that is as simple as collocated cokriging. An intrinsic model of coregionalization is adopted with secondary data at the location being considered and at the locations of the primary data. The resulting technique is referred to as intrinsic collocated cokriging (ICCK).

5.2. Simple Collocated Cokriging is not an Intrinsic Model

It is interesting to note that simple cokriging with the intrinsic correlation model does not reduce to collocated simple cokriging. Let us review a proof of this fact (see also Wackernagel, 2003).

Assume that $Z(\mathbf{u})$ and $Y(\mathbf{u})$ are intrinsically correlated with unit variances. Consider simple cokriging to find the unknown value of the variable of Z at location $\mathbf{u}_0$ based on the neighbor data $Z(\mathbf{u}_\alpha)$ at n sample locations $\mathbf{u}_\alpha$, the corresponding values $Y(\mathbf{u}_\alpha)$ at the same locations and the value of the auxiliary variable $Y(\mathbf{u})$ at the estimation location $\mathbf{u}_0$. Then, the value of the variable of $Z(\mathbf{u})$ at location $\mathbf{u}_0$ is given by the simple cokriging approach as:

$$Z^*_{CSK}(\mathbf{u}_0) = m_Z + \lambda_{Y,0}(Y(\mathbf{u}_0) - m_Y) + \sum_{\alpha=1}^{n} \lambda_{Y,\alpha}(Y(\mathbf{u}_\alpha) - m_Y) + \sum_{\alpha=1}^{n} \lambda_{Z,\alpha}(Z(\mathbf{u}_\alpha) - m_Z), \tag{5.1}$$

where the CSK weights $[\boldsymbol{\lambda}_Z^T \ \boldsymbol{\lambda}_Y^T \ \lambda_{Y,0}]^T$ are given by:

$$\begin{pmatrix} \mathbf{R}_Z & \rho_{YZ}\mathbf{R}_Z & \rho_{YZ}\mathbf{r}_Z \\ \rho_{YZ}\mathbf{R}_Z & \mathbf{R}_Z & \mathbf{r}_Z \\ \rho_{YZ}\mathbf{r}_Z^T & \mathbf{r}_Z^T & 1 \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}_Z \\ \boldsymbol{\lambda}_Y \\ \lambda_{Y,0} \end{pmatrix} = \begin{pmatrix} \mathbf{r}_Z \\ \rho_{YZ}\mathbf{r}_Z \\ \rho_{YZ} \end{pmatrix}, \tag{5.2}$$

where $\mathbf{r}_Z$ is the vector of spatial correlations $\rho(\mathbf{u}_\alpha - \mathbf{u}_0)$, and $\mathbf{R}_Z$ is the matrix of spatial correlations $\rho(\mathbf{u}_\alpha - \mathbf{u}_\beta)$, $\alpha, \beta = 1, \ldots, n$.

Now note that in order for simple cokriging to be reduced to collocated simple cokriging, the vector of weights $[\boldsymbol{\lambda}_Z^T \ \mathbf{0}^T \ \lambda_{Y,0}]^T$ must be the solution of the System 5.2. To check this, substitute $[\boldsymbol{\lambda}_Z^T \ \mathbf{0}^T \ \lambda_{Y,0}]^T$ in System 5.2, then we will obtain

$$\begin{pmatrix} \mathbf{R}_Z & \rho_{YZ}\mathbf{R}_Z & \rho_{YZ}\mathbf{r}_Z \\ \rho_{YZ}\mathbf{R}_Z & \mathbf{R}_Z & \mathbf{r}_Z \\ \rho_{YZ}\mathbf{r}_Z^T & \mathbf{r}_Z^T & 1 \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}_Z \\ \mathbf{0} \\ \lambda_{Y,0} \end{pmatrix} = \begin{pmatrix} \mathbf{r}_Z \\ \rho_{YZ}\mathbf{r}_Z \\ \rho_{YZ} \end{pmatrix}. \tag{5.3}$$

With respect to the value of ρ_{YZ}, we can consider now 3 cases:

1. $\rho_{YZ} = 0$, then we have reduction to simple kriging and $\lambda_{Y,0} = 0$. However, $\lambda_{Y,0} = 0$ cannot be a solution of the collocated cokriging; this is because an

implicit assumption of the collocated cokriging is that the weight assigned to collocated secondary data is not zero (Wackernagel, 2003).

2. $\rho_{YZ} = \pm 1$. Consider first $\rho_{YZ} = 1$, then System 5.3 can be rewritten as

$$\begin{cases} \boldsymbol{\lambda}_Z \mathbf{R}_Z + \lambda_{Y,0}\mathbf{r}_Z = \mathbf{r}_Z \\ \boldsymbol{\lambda}_Z \mathbf{R}_Z + \lambda_{Y,0}\mathbf{r}_Z = \mathbf{r}_Z . \\ \boldsymbol{\lambda}_Z \mathbf{r}_Z^T + \lambda_{Y,0} = 1 \end{cases} \tag{5.4}$$

Looking at System 5.4, it becomes clear that the 1st and 2nd equation in this System are the same, thus System (5.4) can be reduced to collocated simple cokriging system

$$\begin{cases} \boldsymbol{\lambda}_Z \mathbf{R}_Z + \lambda_{Y,0}\mathbf{r}_Z = \mathbf{r}_Z \\ \boldsymbol{\lambda}_Z \mathbf{r}_Z^T + \lambda_{Y,0} = 1 \end{cases} . \tag{5.5}$$

with solution $\boldsymbol{\lambda}_Z = \mathbf{0}$, $\lambda_{Y,0} = 1$. The proof that System 5.3 reduces to the trivial collocated simple cokriging System with solution $\boldsymbol{\lambda}_Z = \mathbf{0}$, $\lambda_{Y,0} = -1$, for $\rho_{YZ} = -1$, can be obtained following the same approach.

It is worth noting that, in general, when the primary and secondary variables are perfectly correlated then the collocated secondary datum (if $\rho_{YZ} = 1$) or negative value of collocated secondary datum (if $\rho_{YZ} = -1$) is considered as primary datum and cokriging reduces to kriging.

3. $\rho_{YZ} \neq 0, \pm 1$, then if we multiply the first equation of matrix 5.3 by ρ_{YZ}, and subtract the result from the second equation in 5.3, we will obtain

$$\boldsymbol{\lambda}_Z \rho_{YZ} \mathbf{R}_Z + \lambda_{Y,0}\mathbf{r}_Z - \rho_{YZ}(\boldsymbol{\lambda}_Z \mathbf{R}_Z + \lambda_{Y,0}\rho_{YZ}\mathbf{r}_Z) = \rho_{YZ}\mathbf{r}_Z - \rho_{YZ}\mathbf{r}_Z ,$$

or

$$\lambda_{Y,0}\mathbf{r}_Z - \lambda_{Y,0}\rho_{YZ}^2\mathbf{r}_Z = \mathbf{0} .$$

Thus,

$$\lambda_{Y,0}(1 - \rho_{YZ}^2)\mathbf{r}_Z = \mathbf{0}. \tag{5.6}$$

Due to the fact that $\rho_{YZ} \neq \pm 1$, and there exists non-zero spatial correlations $\rho(\mathbf{u}_\alpha - \mathbf{u}_\beta), \alpha, \beta = 1, \ldots, n$, we can conclude that necessarily $\lambda_{Y,0} = 0$. However, $\lambda_{Y,0} = 0$ cannot be a solution of the collocated cokriging; this is because an implicit assumption of the collocated cokriging is that the weight assigned to collocated secondary data is not zero (Wackernagel, 2003).

Since neither of the values for ρ_{YZ} yields a reduction of the simple cokriging solution to the nontrivial collocated simple cokriging solution, we conclude that simple cokriging with intrinsic model cannot be reduced to collocated simple cokriging, thus *there is no theoretical justification for selecting only one auxiliary sample (collocated) for cokriging in the case of an intrinsic correlation model.*

5.3. Sources of Variance Inflation in Collocated Cokriging

Sequential Gaussian simulation is based on decomposing the multivariate distribution into a series of conditional distributions for each location (Isaaks, 1990). Simulation is performed by drawing from local conditional distributions defined by collocated cokriging. Newly simulated data are used as conditional data in simulation of subsequent nodes. Multiple equally-probable realizations of the property of interest are created.

A problem of variance inflation is observed in simulation based on collocated simple cokriging. The aim of this Section is to determine the reason for this variance inflation.

Let us consider estimation of the value of the primary standard normal random variable Z at location $\mathbf{u}_0$ using two conditioning (original) data $Z(\mathbf{u}_1)$ and $Z(\mathbf{u}_2)$, the simulated value of the same type at location $\mathbf{u}_3$, $Z(\mathbf{u}_3)$ given by:

$$Z(\mathbf{u}_3) = \lambda_{Y,3} Y(\mathbf{u}_3) + \sum_{\alpha=1}^{2} \lambda_{Z,\alpha} Z(\mathbf{u}_\alpha) + R(\mathbf{u}_3), \tag{5.7}$$

where $Y(\mathbf{u}_3)$ is collocated value to $Z(\mathbf{u}_3)$ of standard normal auxiliary variable, the collocated simple kriging weights $[\lambda_{Z,1}\ \lambda_{Z,2}\ \lambda_{Y,3}]^T$ are given by:

$$\begin{pmatrix} 1 & \rho_{ZZ}(\mathbf{u}_1 - \mathbf{u}_2) & \rho_{YZ}(\mathbf{u}_1 - \mathbf{u}_3) \\ \rho_{ZZ}(\mathbf{u}_2 - \mathbf{u}_1) & 1 & \rho_{YZ}(\mathbf{u}_2 - \mathbf{u}_3) \\ \rho_{YZ}(\mathbf{u}_3 - \mathbf{u}_1) & \rho_{YZ}(\mathbf{u}_3 - \mathbf{u}_2) & 1 \end{pmatrix} \begin{pmatrix} \lambda_{Z,1} \\ \lambda_{Z,2} \\ \lambda_{Y,3} \end{pmatrix} = \begin{pmatrix} \rho_{ZZ}(\mathbf{u}_1 - \mathbf{u}_3) \\ \rho_{ZZ}(\mathbf{u}_2 - \mathbf{u}_3) \\ \rho_{YZ}(\mathbf{0}) \end{pmatrix}, \tag{5.8}$$

or, in System format as

$$\begin{cases} \lambda_{Z,1} + \lambda_{Z,2}\rho_{ZZ}(\mathbf{u}_1 - \mathbf{u}_2) + \lambda_{Y,0}\rho_{YZ}(\mathbf{u}_1 - \mathbf{u}_3) = \rho_{ZZ}(\mathbf{u}_1 - \mathbf{u}_3) \\ \lambda_{Z,1}\rho_{ZZ}(\mathbf{u}_2 - \mathbf{u}_1) + \lambda_{Z,2} + \lambda_{Y,0}\rho_{YZ}(\mathbf{u}_2 - \mathbf{u}_3) = \rho_{ZZ}(\mathbf{u}_2 - \mathbf{u}_3) \\ \lambda_{Z,1}\rho_{YZ}(\mathbf{u}_3 - \mathbf{u}_1) + \lambda_{Z,2}\rho_{YZ}(\mathbf{u}_3 - \mathbf{u}_2) + \lambda_{Y,3} = \rho_{YZ}(\mathbf{0}) \end{cases} \tag{5.9}$$

and $R(\mathbf{u}_3)$ is independent normal random error with mean of zero and variance of $Var(R(\mathbf{u}_3)) = 1 - \lambda_{Z,1}\rho_{ZZ}(\mathbf{u}_1 - \mathbf{u}_3) - \lambda_{Z,2}\rho_{ZZ}(\mathbf{u}_2 - \mathbf{u}_3) - \lambda_{Y,3}\rho_{YZ}(\mathbf{0})$.

The collocated simple cokriging estimate of $Z_{CCSK}{*}(\mathbf{u}_0)$ is given by:

$$Z{*}_{CCSK}(\mathbf{u}_0) = \lambda_{Y,0}Y(\mathbf{u}_0) + \sum_{\alpha=1}^{2}\lambda_{Z,\alpha}Z(\mathbf{u}_\alpha) + \lambda_{Z,3}Z(\mathbf{u}_3), \tag{5.10}$$

where the collocated simple cokriging weights $[\lambda_{Z,1}\ \lambda_{Z,2}\ \lambda_{Z,3}\ \lambda_{Y,0}]^T$ are given by:

$$\begin{cases} \lambda_{Z,1} + \lambda_{Z,2}\rho_{ZZ}(\mathbf{u}_1 - \mathbf{u}_2) + \lambda_{Z,3}\rho_{ZZ}(\mathbf{u}_1 - \mathbf{u}_3) + \lambda_{Y,0}\rho_{YZ}(\mathbf{u}_1 - \mathbf{u}_0) = \rho_{ZZ}(\mathbf{u}_1 - \mathbf{u}_0) \\ \lambda_{Z,1}\rho_{ZZ}(\mathbf{u}_2 - \mathbf{u}_1) + \lambda_{Z,2} + \lambda_{Z,3}\rho_{ZZ}(\mathbf{u}_2 - \mathbf{u}_3) + \lambda_{Y,0}\rho_{YZ}(\mathbf{u}_2 - \mathbf{u}_0) = \rho_{ZZ}(\mathbf{u}_2 - \mathbf{u}_0). \\ \lambda_{Z,1}\rho_{ZZ}(\mathbf{u}_3 - \mathbf{u}_1) + \lambda_{Z,2}\rho_{ZZ}(\mathbf{u}_3 - \mathbf{u}_2) + \lambda_{Z,3} + \lambda_{Y,0}\rho_{YZ}(\mathbf{u}_3 - \mathbf{u}_0) = \rho_{ZZ}(\mathbf{u}_3 - \mathbf{u}_0) \\ \lambda_{Z,1}\rho_{YZ}(\mathbf{u}_0 - \mathbf{u}_1) + \lambda_{Z,2}\rho_{YZ}(\mathbf{u}_0 - \mathbf{u}_2) + \lambda_{Z,3}\rho_{YZ}(\mathbf{u}_0 - \mathbf{u}_3) + \lambda_{Y,0} = \rho_{YZ}(\mathbf{0}) \end{cases} \tag{5.11}$$

Let us note that for any i =1, 2, where i stands for the number of conditioning (original) neighbor data, the following holds:

$$\begin{aligned} & Cov(Z(\mathbf{u}_i), Z{*}_{CCSK}(\mathbf{u}_0)) \\ & = Cov\left(\lambda_{Y,0}Y(\mathbf{u}_0) + \sum_{\alpha=1}^{2}\lambda_{Z,\alpha}Z(\mathbf{u}_\alpha) + \lambda_{Z,3}Z(\mathbf{u}_3), Z(\mathbf{u}_i)\right) \\ & = \lambda_{Y,0}Cov(Y(\mathbf{u}_0), Z(\mathbf{u}_i)) + \lambda_{Z,1}Cov(Z(\mathbf{u}_1), Z(\mathbf{u}_i)) + \lambda_{Z,2}Cov(Z(\mathbf{u}_2), Z(\mathbf{u}_i)) + \lambda_{Z,3}Cov(Z(\mathbf{u}_3), Z(\mathbf{u}_i)) \\ & = \lambda_{Y,0}\rho_{YZ}(\mathbf{u}_i - \mathbf{u}_0) + \lambda_{Z,1}\rho_{ZZ}(\mathbf{u}_1 - \mathbf{u}_i) + \lambda_{Z,2}\rho_{ZZ}(\mathbf{u}_2 - \mathbf{u}_i) + \lambda_{Z,3}\rho_{ZZ}(\mathbf{u}_i - \mathbf{u}_3) = \rho_{ZZ}(\mathbf{u}_i - \mathbf{u}_0). \end{aligned}$$

Thus, the covariance (correlation) between the new estimate and the conditioning data values is correct. Note that the last two substitutions in the above derivation followed from Systems (5.9) and (5.11), respectively. Moreover, note that

$$\begin{aligned} & Cov(Z(\mathbf{u}_3), Z{*}_{CCSK}(\mathbf{u}_0)) \\ & = Cov\left(Z(\mathbf{u}_3), \lambda_{Y,0}Y(\mathbf{u}_0)) + \sum_{\alpha=1}^{2}\lambda_{Z,\alpha}Z(\mathbf{u}_\alpha) + \lambda_{Z,3}Z(\mathbf{u}_3)\right) \\ & = \lambda_{Y,0}Cov(Z(\mathbf{u}_3), Y(\mathbf{u}_0)) + \lambda_{Z,1}Cov(Z(\mathbf{u}_3), Z(\mathbf{u}_1)) + \lambda_{Z,2}Cov(Z(\mathbf{u}_3), Z(\mathbf{u}_2)) + \lambda_{Z,3}Var(Z(\mathbf{u}_3)) \\ & = \lambda_{Y,0}Cov(Z(\mathbf{u}_3), Y(\mathbf{u}_0)) + \lambda_{Z,1}\rho_{ZZ}(\mathbf{u}_3 - \mathbf{u}_1) + \lambda_{Z,2}\rho_{ZZ}(\mathbf{u}_3 - \mathbf{u}_2) + \lambda_{Z,3}. \end{aligned}$$

If $Cov(Z(\mathbf{u}_3), Y(\mathbf{u}_0)) = \rho_{YZ}(\mathbf{u}_3 - \mathbf{u}_0)$, then the last substitution from System (5.9) would result in the correct covariance between the new estimate and the previously calculated estimates. However, $Cov(Z(\mathbf{u}_3), Y(\mathbf{u}_0))$ cannot be equal to $\rho_{YZ}(\mathbf{u}_3 - \mathbf{u}_0)$. Let us prove this using the method of contradiction. Specifically, let us assume that $Cov(Z(\mathbf{u}_3), Y(\mathbf{u}_0)) = \rho_{YZ}(\mathbf{u}_3 - \mathbf{u}_0)$, then it follows that the simple cokriging problem would be necessarily reduced to the collocated simple cokriging problem (which inherently ensures this equality), however, as shown above such reduction is impossible in the case of intrinsic correlation model. Therefore $Cov(Z(\mathbf{u}_3), Y(\mathbf{u}_0))$ cannot be equal to $\rho_{YZ}(\mathbf{u}_3 - \mathbf{u}_0)$ and we can conclude that due to the fact that the collocated cokriging System has no reproduction of cross covariance $Cov(Z(\mathbf{u}_3), Y(\mathbf{u}_0))$, the correct covariance between the new estimate and the previously calculated estimates cannot be ensured. When collocated simple cokriging is put into sequential simulation mode, the incorrect covariance between simulated data is translated into biasness of newly simulated data. This is because simulation is sequential. Previously simulated data are used as a conditioning data for calculating new simulated nodes; the simulated data do not have the correct covariance between each other, thus, System (5.9) results in incorrect weights and biased estimates. And, as a result, variance inflation is often observed. Note that variance inflation rather than variance deflation is observed because local conditional distributions (from which simulated values are drawn) have larger variances in collocated cokriging than in intrinsic collocated cokriging.

In view of the above analysis, a natural solution for the problem of variance inflation in sequential simulation is to consider full simple cokriging based on the intrinsic model (intrinsic collocated cokriging) instead of the Markov model.

5.4. Intrinsic Collocated Cokriging: Examples

5.4.1. Example 1

Let us consider the following linear model of coregionalization (LMC) for the primary standard normal random variable *Z* and secondary standard normal random variable *Y*:

$$\begin{aligned} \gamma_{YY}(\mathbf{h}) &= 0.1 \cdot Sph_{16}(\mathbf{h}) + 0.9 \cdot Gaus_{32}(\mathbf{h}) \\ \gamma_{YZ}(\mathbf{h}) &= 0.25 \cdot Sph_{16}(\mathbf{h}) + 0.25 \cdot Gaus_{32}(\mathbf{h}) \ , \\ \gamma_{ZZ}(\mathbf{h}) &= 0.9 \cdot Sph_{16}(\mathbf{h}) + 0.1 \cdot Gaus_{32}(\mathbf{h}) \end{aligned} \tag{5.12}$$

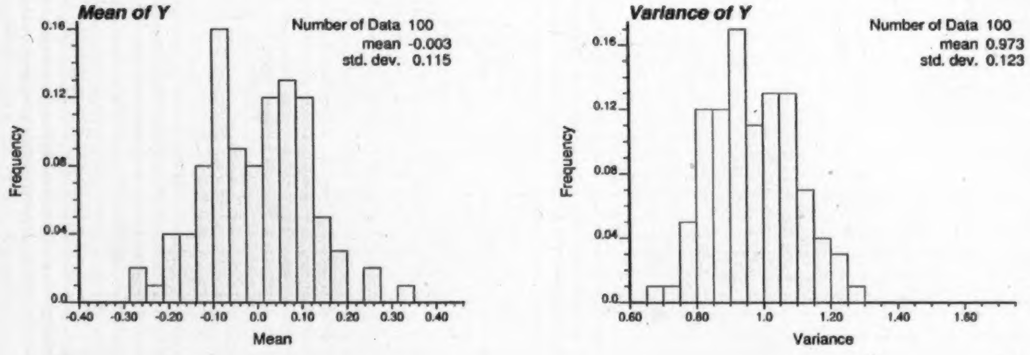
where $Sph_{16}(\mathbf{h}), Gaus_{32}(\mathbf{h})$ denote the isotropic Spherical variogram model with the range of 16 and isotropic Gaussian variogram model with the practical range of 32 (see Chapter 2 for reference on variogram models). Note that System (5.12) is a valid LMC, since

$$0.1 \cdot 0.9 = 0.09 \geq 0.25 \cdot 0.25 = 0.0626 \text{ and } 0.9 \cdot 0.1 = 0.09 \geq 0.25 \cdot 0.25 = 0.0625 .$$

The correlation at lag 0 between the primary and secondary random variables is $\rho_{YZ} = 0.5$. This 2-D example was chosen small enough to easily visualize the results yet large enough to show realistic variations in the results. The conclusions drawn from this example are considered reasonably general.

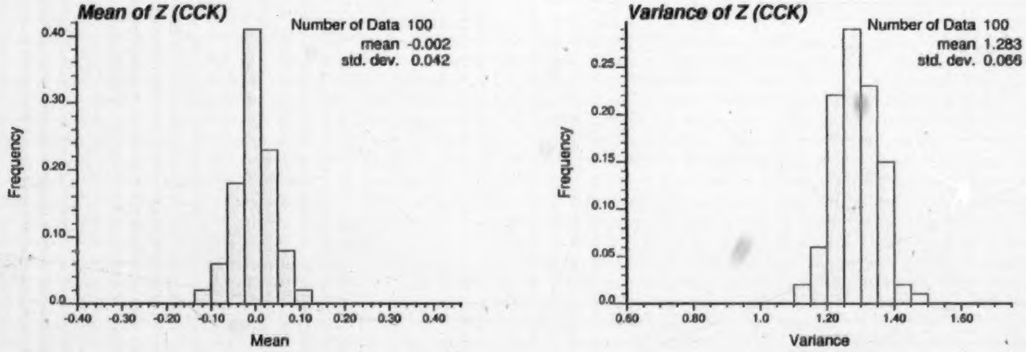
Now let us consider unconditional sequential Gaussian simulation (SGS) based on the simple collocated cokriging for the primary variable *Z* (continuity of Z is given by $\gamma_{ZZ}(\mathbf{h})$ in System 5.12; cross-covariance is build based on Markov model I) using exhaustive secondary random variable *Y* and coefficient of correlation $\rho_{YZ} = 0.5$. The exhaustive secondary information for *Y* was obtained by unconditional sequential Gaussian simulation (SGS) with variogram model $\gamma_{YY}(\mathbf{h})$ given in System 5.12. Figure 5.1 shows the distributions of the means and variances of the secondary random variable for 100 SGS realizations. In sequential simulation for both primary and secondary random variables, the maximum number of simulated nodes to use is set to 12 and the maximum search radii is set to largest variogram range, that is, 32.

A summary of the results for the primary random variable *Z* for 100 SGS realizations of the area of 256 by 256 cells is shown in Figure 5.2.



a) b)

Figure 5.1: Distribution of the means of the secondary standard normal random variable *Y* for the 100 sequential Gaussian realizations (a); and distribution of the variances of the secondary standard normal random variable *Y* for the 100 sequential Gaussian realizations (b).



a) b)

Figure 5.2: Distribution of the means of the primary standard normal random variable *Z* for the 100 sequential Gaussian realizations based on simple collocated cokriging (a); and distribution of the variances of the primary standard normal random variable *Z* for the 100 sequential Gaussian realizations based on simple collocated cokriging (b).

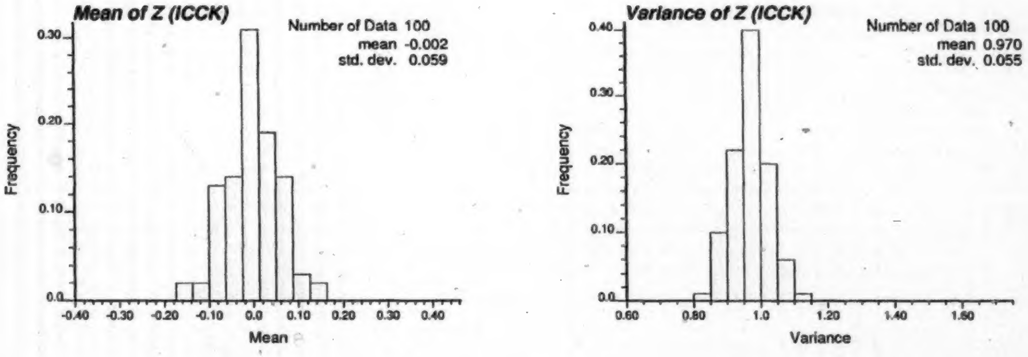
Figure 5.2 shows the distributions of the means and variances of the primary random variable Z realizations obtained based on sequential Gaussian simulation with simple collocated cokriging. Note that the expected mean of the distribution of the primary random variable Z modeled by sequential Gaussian simulation with simple collocated cokriging is virtually zero. There is, however, a dramatic deviation from the target variance of one.

The variance of the primary random variable Z modeled by sequential Gaussian simulation with simple collocated cokriging is, on average, around 28% higher than 1. Note that due to a finite domain size and conditioning data, the target variance reproduced by simulation should be even less than 1. This makes the result of 1.28 obtained in sequential Gaussian simulation with collocated cokriging even worse than it appears. One can argue that this strong deviation is a consequence of the mismatch in the continuity of the primary and secondary random variable (Markov model is inappropriate); however, this is not true. Let us consider modeling of the primary random variable Z using sequential Gaussian simulation with intrinsic collocated cokriging (ICCK). Intrinsic correlation model is built based on the variogram of the primary variable Z, that is, $\gamma_{YY}(\mathbf{h}) = \gamma_{ZZ}(\mathbf{h}); \quad \gamma_{YZ}(\mathbf{h}) = 0.5 \cdot \gamma_{ZZ}(\mathbf{h})$.

The ICCK acronym may be slightly confusing. The proposal is to use more than the secondary datum at the location being estimated. The notion of "collocated" must be extended; secondary data at other locations must be used as well. Then there are two options to explore: (1) secondary data collocated with all primary data, and (2) secondary data in the local neighborhood of the estimation location.

Let us explore option 1 first. Secondary information in intrinsic collocated simple cokriging is selected at the same locations as the primary data and at the estimation location. Note that if the secondary data were selected only at the locations of the available primary data, then the simple cokriging with the intrinsic model of coregionalization would necessarily reduce to simple kriging. However, this screening effect vanishes when a collocated secondary data is added to estimation; the cause of this is a "relay effect" with the other secondary data (Chiles and Delfiner, 1999).

Figure 5.3 shows the distributions of the means and variances of the primary random variable Z obtained in this case.



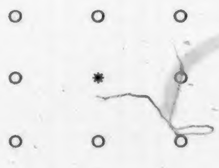
a) b)

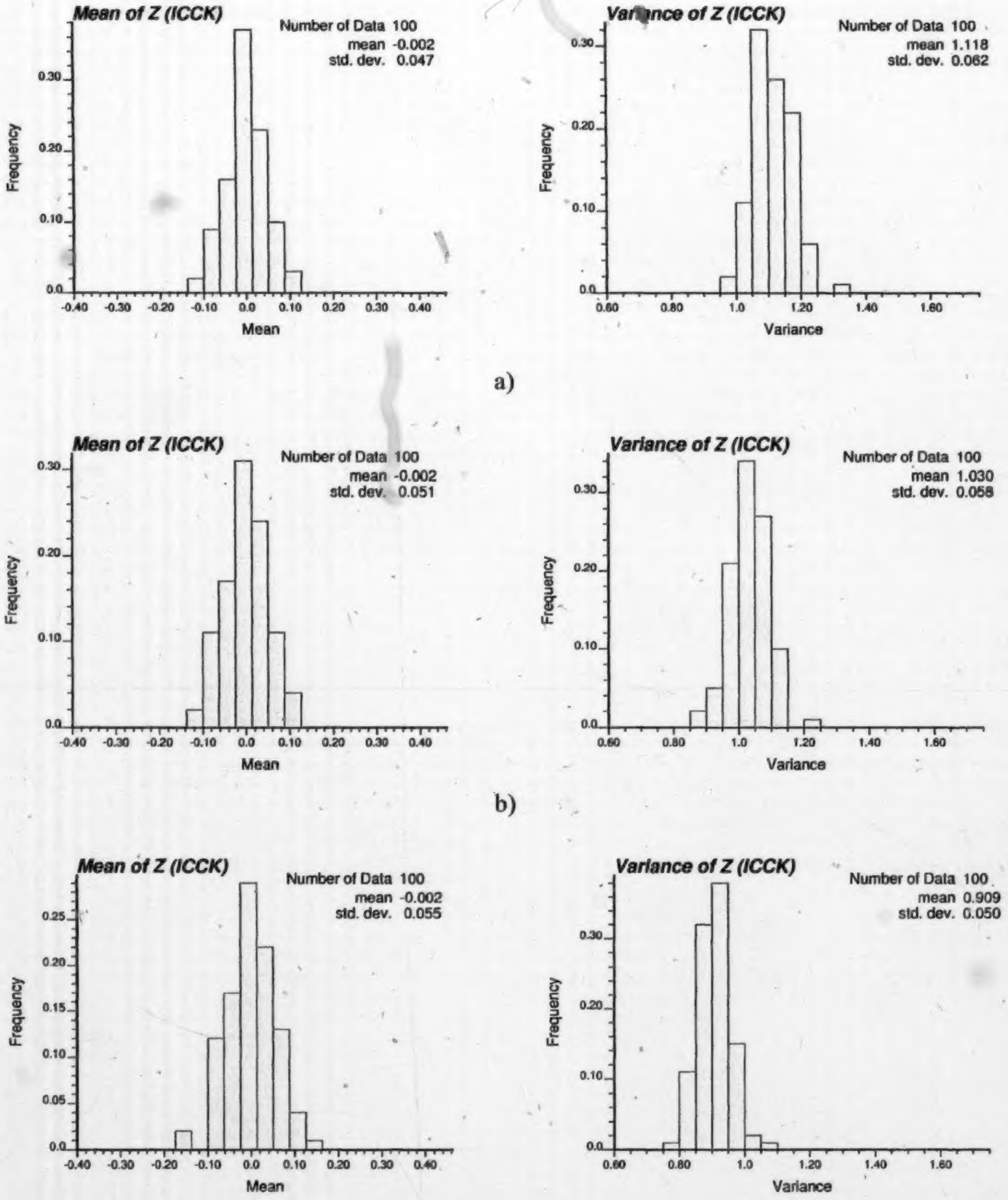
Figure 5.3: Distribution of the means of the primary random variable Z for the 100 sequential Gaussian realizations based on intrinsic collocated cokriging (a); and distribution of the variances of the primary random variable Z for the 100 sequential Gaussian realizations based on intrinsic collocated cokriging (b).

Note from Figure 5.3 that both the expected mean and expected variance are virtually the same as the target mean of 0 and the target variance of 1. Thus, we can see that the main factor which triggers the variance inflation in sequential Gaussian simulation with simple collocated cokriging is not simply the assumption of Markov model, but using only one collocated data when performing cokriging.

Figure 5.4 shows the results for the means and variances of the primary random variable Z obtained using option 2. That is, sequential Gaussian simulation with intrinsic collocated cokriging when the secondary data are selected in the local neighborhood of the estimation location according to the following configuration:

Schematic 1:





a)

b)

c)

Figure 5.4: Distribution of the means (first column) and variances (second column) of the primary random variable Z for the 100 sequential Gaussian realizations based on intrinsic collocated cokriging for the secondary data configuration shown in Schematic 1 with secondary data separation distance in X and Y directions: 1 grid cell (a); 2 grid cells (b) and 5 grid cells (c), respectively.

Clearly, reproduction of the target statistics shown in Figure 5.4 for intrinsic collocated cokriging is much better than the respective results of sequential Gaussian simulation with simple collocated cokriging. However, it can be also seen from Figure 5.4 that reproduction of the target statistics depends on the chosen secondary data configuration in the local neighborhood of the estimation location.

Thus, despite that the secondary information in collocated cokriging with ICCK do not necessarily need to be taken at the same locations as the primary data; our recommendation is to use secondary data at all primary data locations. This is because it is theoretically valid and computationally effective to take the secondary data at the same primary data locations, and moreover, there is no dependence of the results on the chosen secondary data configuration in the local neighborhood of the estimation location.

Another advantage of using the sequential Gaussian simulation with intrinsic collocated cokriging over sequential Gaussian simulation with simple collocated cokriging is the slightly improved primary variable variogram reproduction. Figures 5.5-5.7 show variogram reproduction of the exhaustive secondary random variable *Y* by the unconditional sequential Gaussian simulation; variogram reproduction of the primary random variable *Z* by the unconditional sequential Gaussian simulation with simple collocated cokriging; and variogram reproduction of the primary random variable *Z* by the unconditional sequential Gaussian simulation with intrinsic collocated cokriging, respectively.

It is apparent from Figures 5.6 and 5.7 that the mismatch between target semivariogram for the primary variable Z is reduced by applying intrinsic collocated cokriging. Note that the amount of mismatch could also depend on such parameters as maximum number of nodes used in simulation, search radii, etc. However, in the case study considered in this paper these parameters were fixed. Note also that if the continuity of the primary and secondary random variables are the same, there is no mismatch between the target semivariogram and semivariogram reproduced in the unconditional sequential Gaussian simulation with intrinsic collocated cokriging.

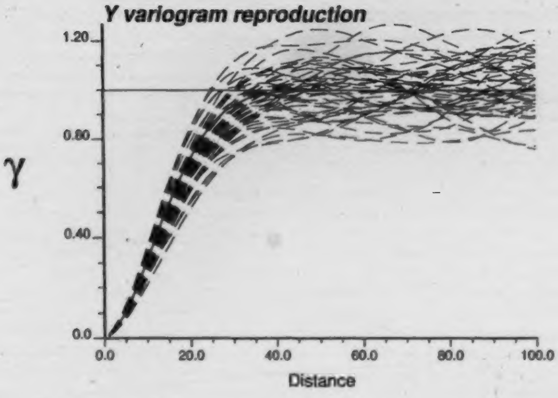


Figure 5.5: Reproduction of the secondary variable *Y* semivariogram by sequential Gaussian simulation.

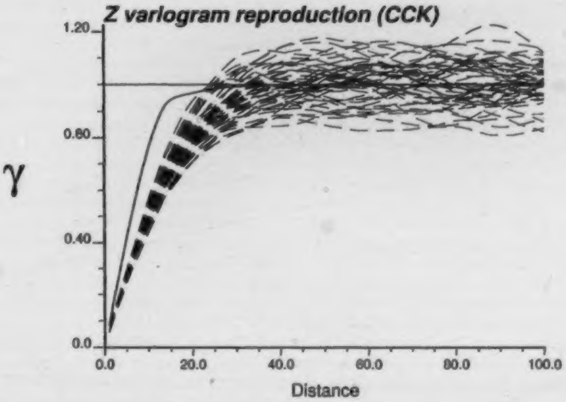


Figure 5.6: Reproduction of the primary variable *Z* semivariogram by the sequential Gaussian simulation with simple collocated cokriging.

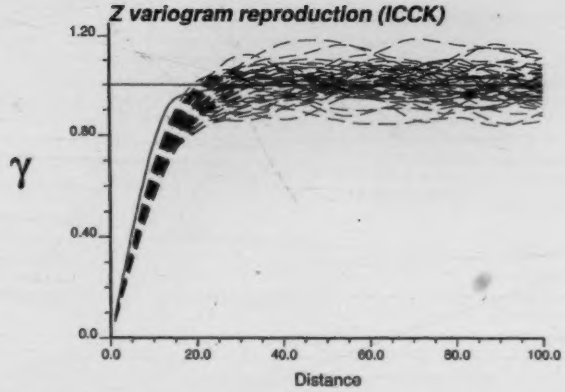


Figure 5.7: Reproduction of the primary variable *Z* semivariogram by sequential Gaussian simulation with intrinsic collocated cokriging.

It is also important to note that modeling of the primary variable based on the secondary random variable in the intrinsic collocated cokriging framework also ensures reproduction of the correlation between primary and secondary random variable. This point is illustrated in Figure 5.8. Figure 5.8 shows 100 correlation coefficients obtained for each of the 100 SGS realizations with intrinsic collocated cokriging. The observed mean correlation coefficient of 0.4640 meets closely the target correlation coefficient of 0.5 used in simulation.

5.4.2. Example 2

The use of real data is problematic due to the cost and confidentiality of seismic data; however, synthetic data are fashioned after a number of real case studies where variance inflation was noted to be a problem.

Figure 5.9 shows locations of 20 primary data in a study domain of size 100 by 100 units. Figure 5.9 also shows the primary data distribution, the crossplot between primary data and collocated secondary data and the map of exhaustive secondary data. All data are in Gaussian units.

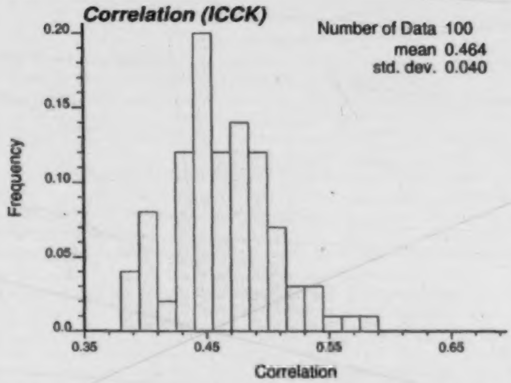


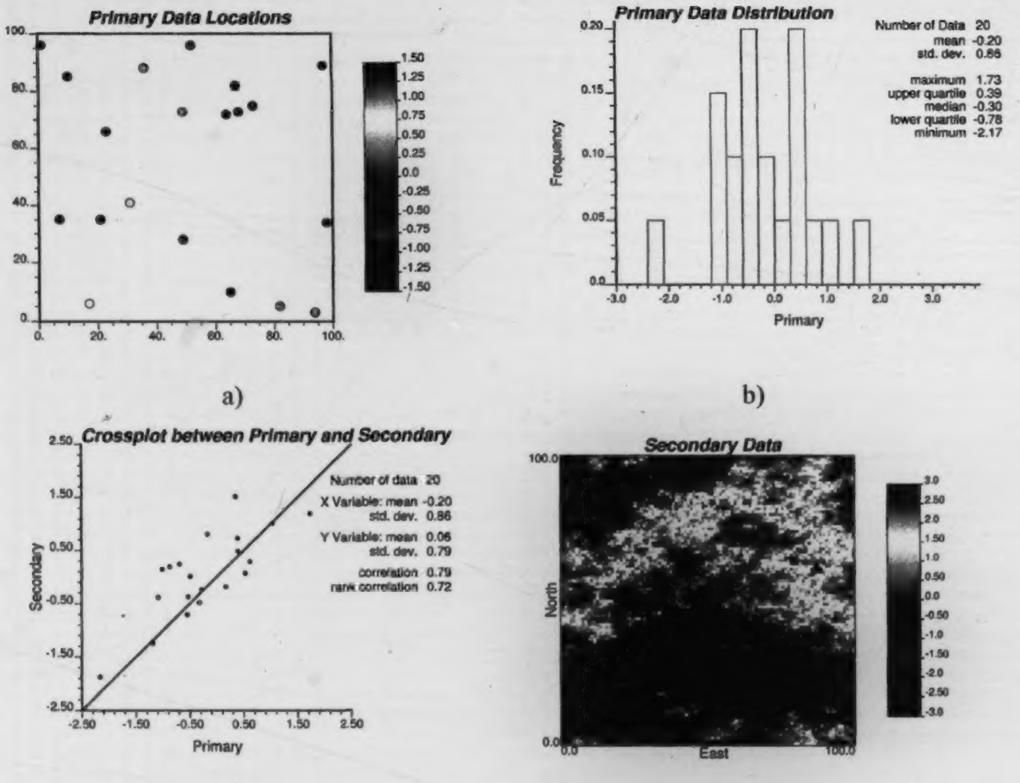
Figure 5.8: Reproduction of the correlation between primary and secondary random variables for sequential Gaussian simulation with intrinsic collocated cokriging.

The following linear model of coregionalization describes the joint continuity of the primary and secondary data:

$$\begin{aligned} \gamma_{YY}(\mathbf{h}) &= 0.3 \cdot Exp_{\substack{a_1=10\\a_2=20}}(\mathbf{h}) + 0.7 \cdot Sph_{\substack{a_1=20\\a_2=40}}(\mathbf{h}) \\ \gamma_{YZ}(\mathbf{h}) &= 0.45 \cdot Exp_{\substack{a_1=10\\a_2=20}}(\mathbf{h}) + 0.35 \cdot Sph_{\substack{a_1=20\\a_2=40}}(\mathbf{h}), \\ \gamma_{ZZ}(\mathbf{h}) &= 0.8 \cdot Exp_{\substack{a_1=10\\a_2=20}}(\mathbf{h}) + 0.2 \cdot Sph_{\substack{a_1=20\\a_2=40}}(\mathbf{h}) \end{aligned} \tag{5.15}$$

Now, let us compare reproduction of the target mean and variance statistics by sequential Gaussian simulation (SGS) based on simple collocated cokriging and by sequential Gaussian simulation (SGS) based on intrinsic collocated cokriging for the primary variable of Figure 5.9. Example realizations obtained by both methods are shown in Figure 5.10. Figure 5.10 shows the similarity in the results; however, note that the realization obtained with intrinsic collocated cokriging contain less extreme values than the realization obtained based on simple collocated cokriging.

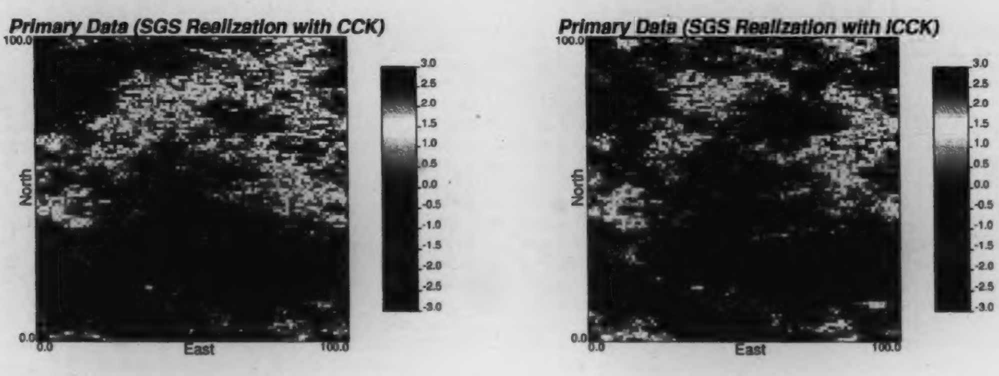
Figure 5.11 shows the results for the mean and variance of 100 SGS realization of the primary variable over the area of 100 by 100 units obtained based on simple collocated cokriging. Results obtained based with intrinsic collocated cokriging are shown in Figure 5.12.



a) b)

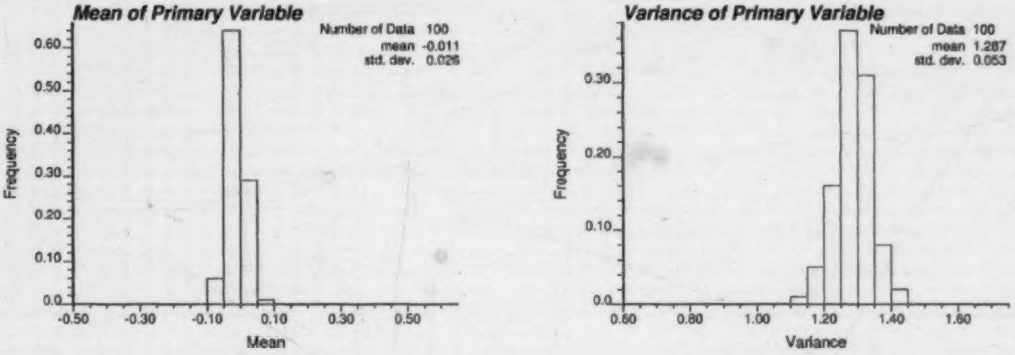
c) d)

Figure 5.9: Locations of the 20 primary data (a) and their distribution (b); the crossplot between primary data and collocated secondary data (c) and the map of exhaustive secondary data (d). The data are in Gaussian units.


Primary Data (SGS Realization with CCK)
100.0
North
0.0
0.0
East
100.0
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
Primary Data (SGS Realization with ICCK)
100.0
North
0.0
0.0
East
100.0
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0

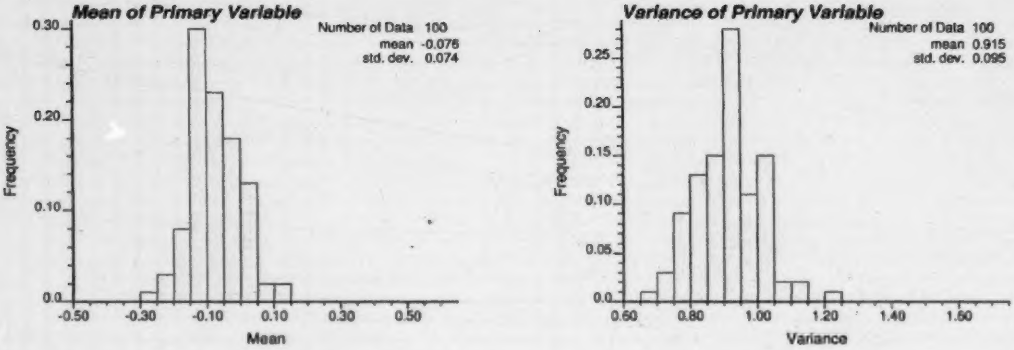
a) b)

Figure 5.10: Example sequential Gaussian realization obtained based on simple collocated cokriging (a) and intrinsic collocated cokriging (b).



a) b)

Figure 5.11: Distribution of the means of the primary random variable for the 100 sequential Gaussian realizations based on simple collocated cokriging (a); and distribution of the variances of the primary random variable for the 100 sequential Gaussian realizations based on simple collocated cokriging (b).



a) b)

Figure 5.12: Distribution of the means of the primary random variable for the 100 sequential Gaussian realizations based on intrinsic collocated cokriging (a); and distribution of the variances of the primary random variable for the 100 sequential Gaussian realizations based on intrinsic collocated cokriging (b).

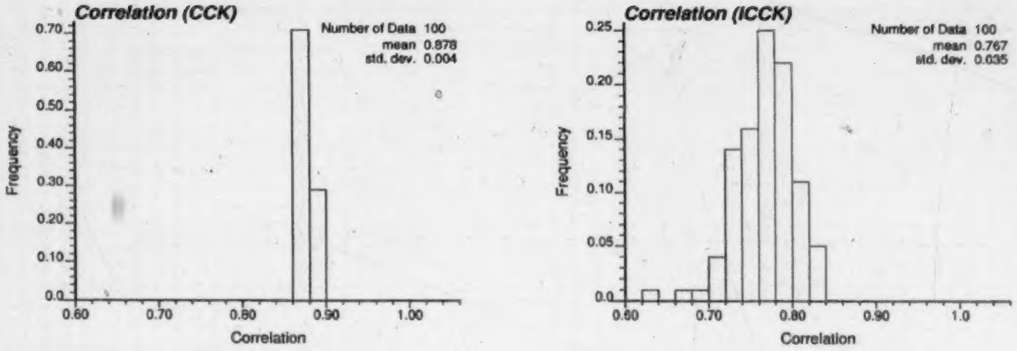
Note that intrinsic collocated cokriging used in sequential mode does not result in variance inflation. Moreover, from Figure 5.13 we can also note that the target correlation of 0.8 is much better reproduced by intrinsic collocated cokriging, compared to SGS with simple collocated cokriging.

5.5. Cokriging versus Collocated Cokriging

This example compares the results to cokriging with a linear model of coregionalization. Let us consider the same as before linear model of coregionalization for the primary standard normal random variable Z and secondary standard normal random variable Y

$$\begin{aligned} \gamma_{YY}(h) &= 0.1 \cdot Sph_{16}(h) + 0.9 \cdot Gaus_{32}(h) \\ \gamma_{YZ}(h) &= 0.25 \cdot Sph_{16}(h) + 0.25 \cdot Gaus_{32}(h) \ , \\ \gamma_{ZZ}(h) &= 0.9 \cdot Sph_{16}(h) + 0.1 \cdot Gaus_{32}(h) \end{aligned} \tag{5.16}$$

where $Sph_{16}(h), Gaus_{32}(h)$ denote the Spherical variogram model with the range of 16 and Gaussian variogram model with the range of 32.



a) b)

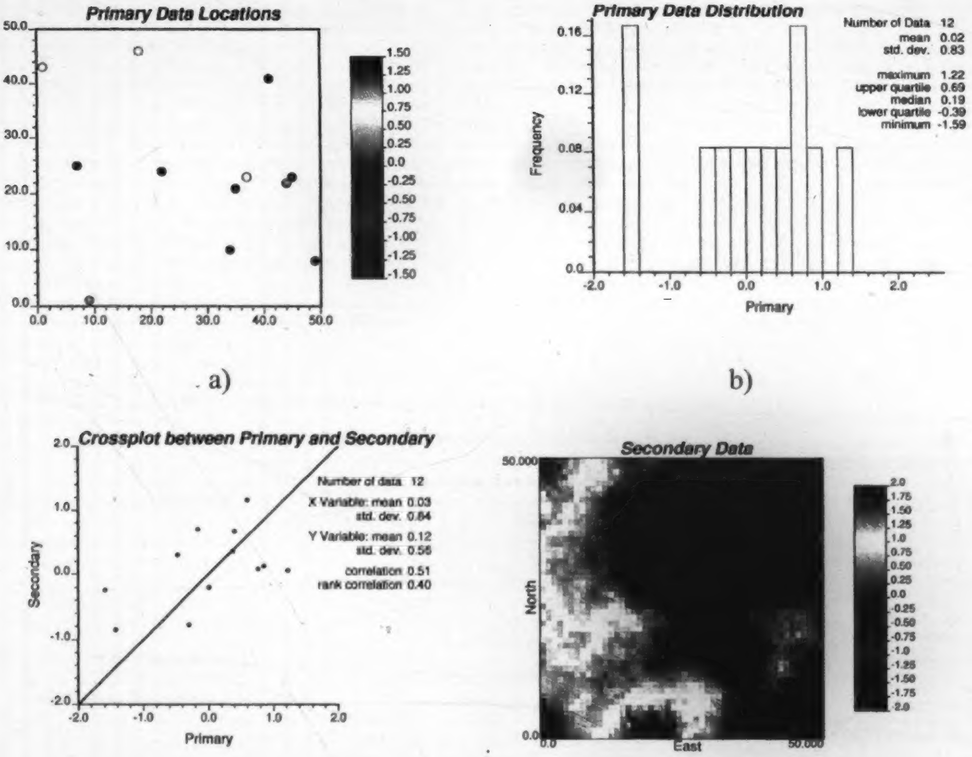
Figure 5.13: Reproduction of the correlation between primary and secondary random variables for sequential Gaussian simulation with simple collocated cokriging (a) and sequential Gaussian simulation with intrinsic collocated cokriging (b).

Now let us consider estimation of the domain 50 by 50 units based on the primary data and exhaustive secondary data. Figure 5.14 shows locations of 12 primary data and their distribution, the crossplot between primary data and collocated secondary data and the map of exhaustive secondary data. Three approaches for estimation are considered:

- simple cokriging with the linear model of corregionalization (5.16);
- simple cokriging with the intrinsic correlation model. The primary variable variogram of System (5.16) is taken as the underlying variogram model for the intrinsic model of coregionalization;
- And, finally, using collocated simple cokriging.

5.5.1. Difference in the Profiles of Weights

Let us first perform estimation of the two arbitrary locations in the study domain, say (10,10) and (35, 35), based on simple cokriging, collocated simple cokriging and intrinsic collocated cokriging and analyze the difference in the profiles of weights.



a) b)

c) d)

Figure 5.14: Locations of 12 primary data (a) and their distribution (b), the crossplot between primary data and collocated secondary data (c) and the map of exhaustive secondary data (d).

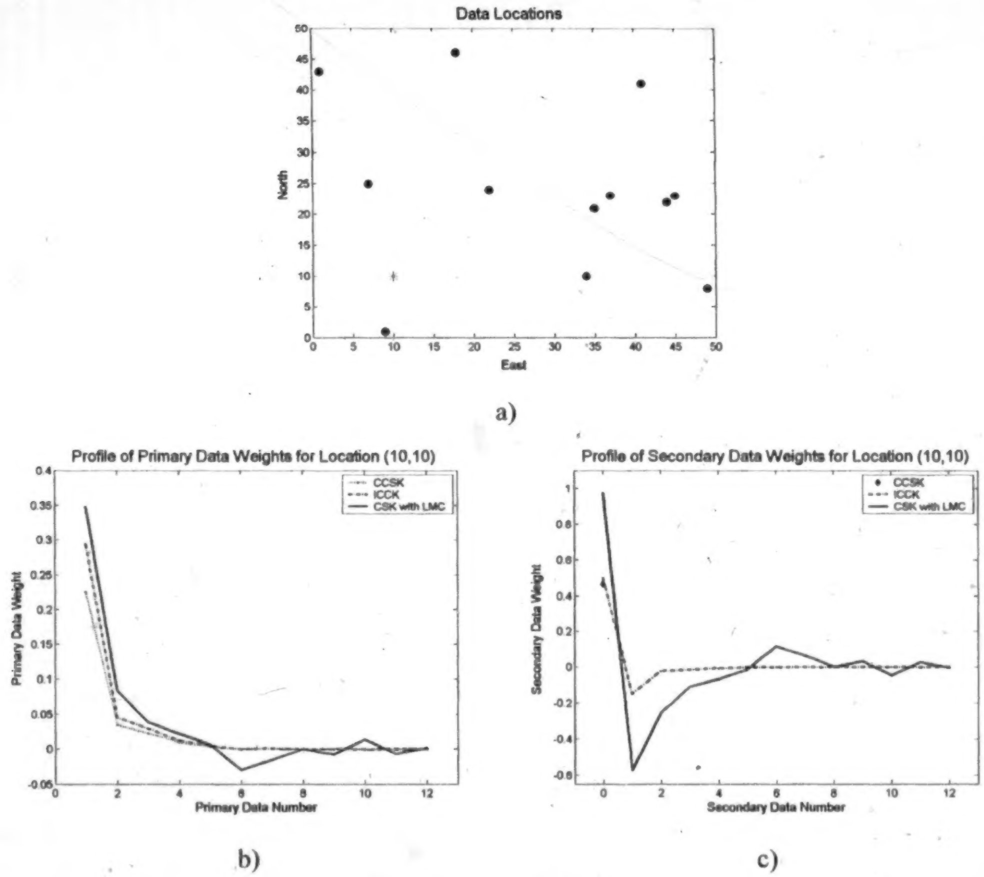
Figures 5.15 and 5.16 show for each of the two locations of interest the estimation variances, accumulated weights and the weights profiles. Note that accumulated primary and secondary data weights are calculated as a sum of all weights given to all primary and all secondary data, respectively. It is interesting to note from Figures 5.15 and 5.16 that intrinsic collocated cokriging assigns the collocated secondary data weight equal to the correlation coefficient between the primary and secondary data; the largest weight assigned to the collocated secondary data is obtained in simple cokriging with linear model of coregionalization (LMC) and the smallest in the collocated simple cokriging. Note, however, that despite the LMC assigns the largest weight to the collocated data, the accumulated weight assigned to all secondary data in simple cokriging with linear model of coregionalization is the smallest, the largest accumulated weight assigned to secondary data, that is one which is collocated, is obtained in collocated simple cokriging.

The largest accumulated weight assigned to the primary data is obtained in simple cokriging with linear model of coregionalization; the smallest is obtained in collocated simple cokriging.

Intrinsic collocated cokriging method provides an intermediate case in-between the other two correlation models also in terms of estimation variance. The smallest estimation variance is obtained in simple cokriging with LMC. This is, of course, because the simple cokriging is theoretically optimal; that is provides estimates with smallest estimation variance. The largest estimation variance is obtained in collocated simple cokriging.

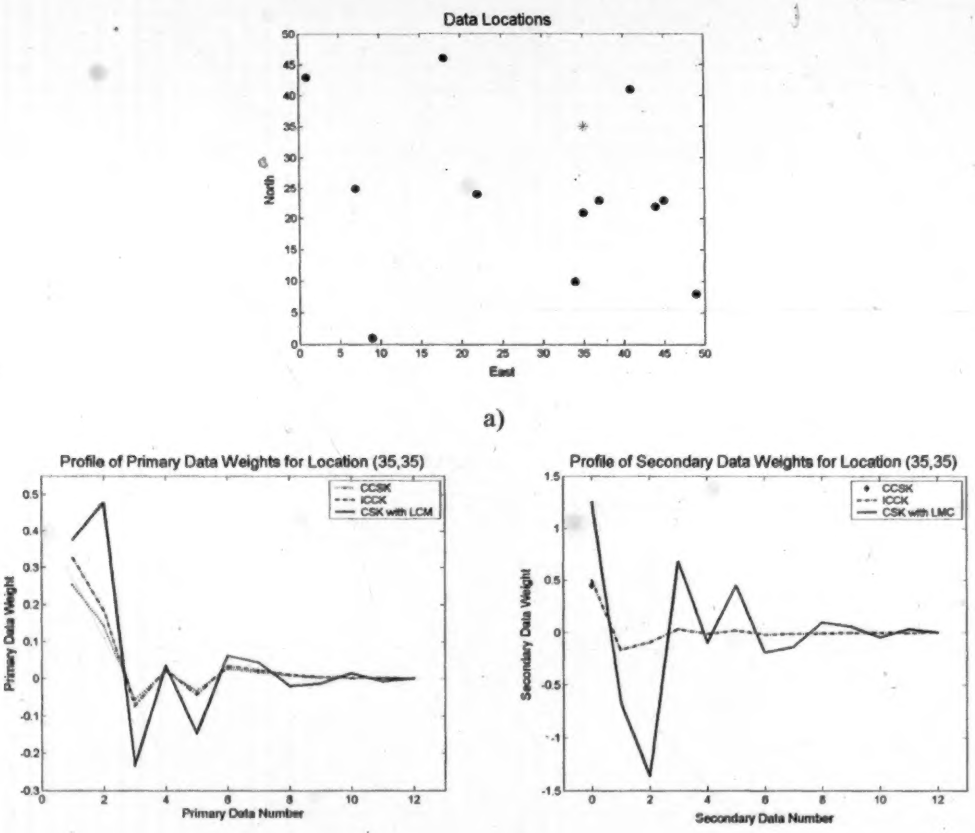
5.5.2. Difference in the Results of Estimation

Now let us consider estimation of the entire domain of study. Figure 5.17 shows the maps of the estimates (means of the local conditional distributions) and estimation variances (variances of the local conditional distributions) obtained based on simple cokriging with LMC, collocated simple cokriging and intrinsic collocated simple cokriging.



	Sum of Primary Variable Weights	Sum of Secondary Variable Weights	Weight received by Collocated Data	Sum of Cokriging Weights	Estimate	Estimation Variance
CCSK	0.2974	0.4650	0.4650	0.7625	0.2528	0.6976
CSK (Intrinsic)	0.3875	0.3062	0.5000	0.6938	0.1735	0.6817
CSK (LMC)	0.4511	0.1246	0.9715	0.5757	-0.0331	0.5961

Figure 5.15: Study domain with conditioning data (circles) and the estimation location (10, 10) (asterisk) (a); Primary data weights as a function of the ordered conditioning data, ordered according to the closeness to the estimation location (b) and secondary data weights as a function of the ordered conditioning data, zero stands for the estimation location (c).



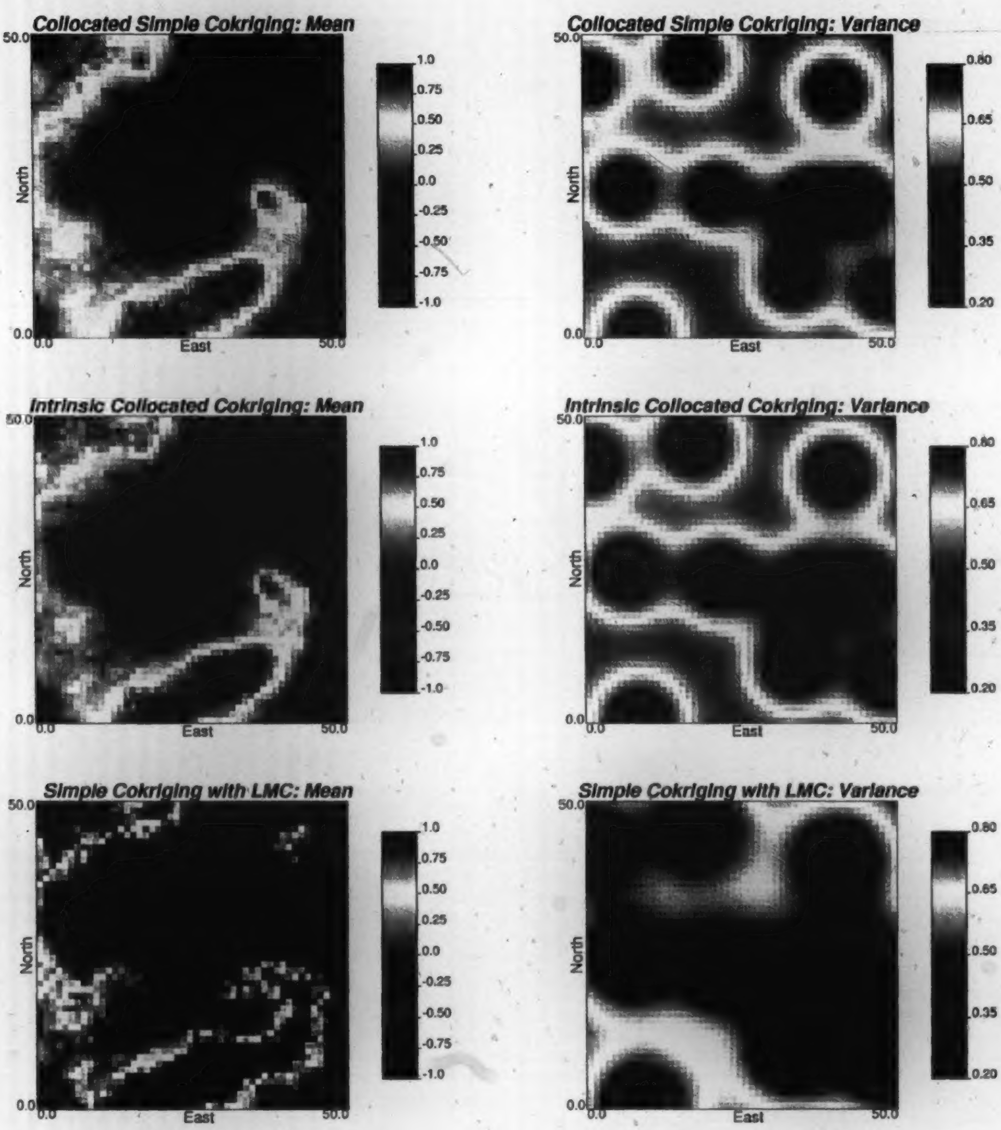
a)

b)

c)

	Sum of Primary Variable Weights	Sum of Secondary Variable Weights	Weight received by Collocated Data	Sum of Cokriging Weights	Estimate	Estimation Variance
CCSK	0.3824	0.4497	0.4497	0.8321	-0.5751	0.6746
CSK (Intrinsic)	0.4933	0.2534	0.5000	0.7466	-0.5270	0.6527
CSK (LMC)	0.5789	0.0683	1.2575	0.6472	-1.0345	0.5153

Figure 5.16: Study domain with conditioning data (circles) and the estimation location (35, 35) (asterisk) (a); Primary data weights as a function of the ordered conditioning data, ordered according to the closeness to the estimation location (b) and secondary data weights as a function of the ordered conditioning data, zero stands for the estimation location (c).


Collocated Simple Cokriging: Mean
Collocated Simple Cokriging: Variance
Intrinsic Collocated Cokriging: Mean
Intrinsic Collocated Cokriging: Variance
Simple Cokriging with LMC: Mean
Simple Cokriging with LMC: Variance
North
East
0.0
50.0
1.0
0.75
0.50
0.25
0.0
-0.25
-0.50
-0.75
-1.0
0.80
0.65
0.50
0.35
0.20

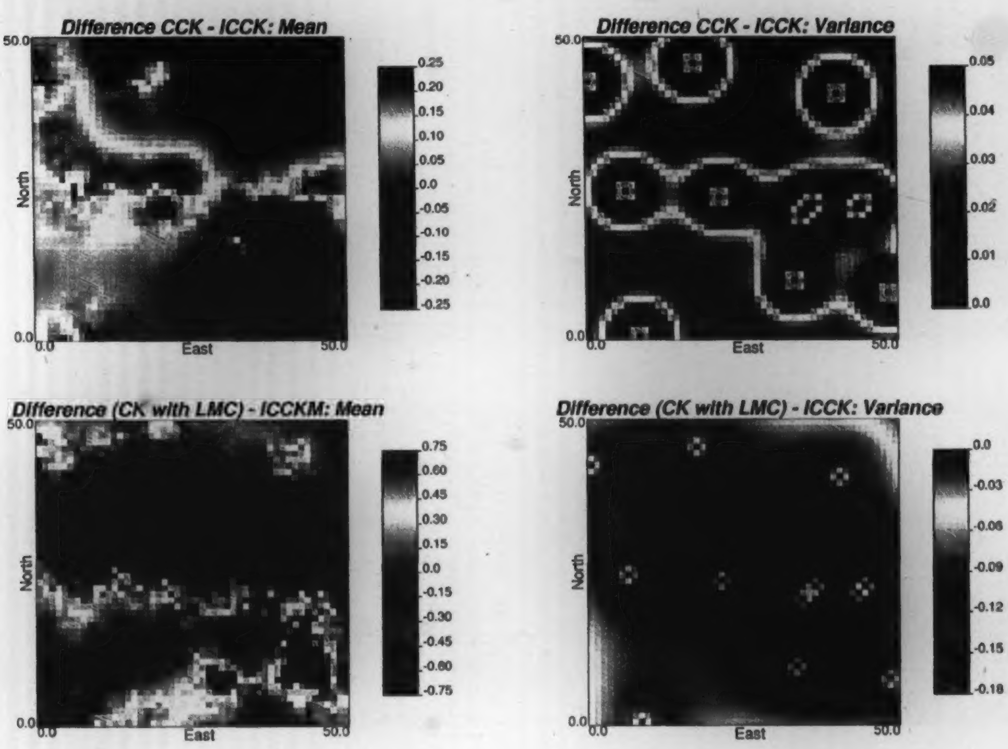
a) b)

Figure 5.17: The maps of estimates (a) and estimation variances (b) obtained based on collocated simple cokriging (top), simple cokriging with the intrinsic correlation model (middle) and simple cokriging with the linear model of corregionalization (bottom).

From Figure 5.17 note that collocated simple cokriging and intrinsic collocated cokriging are very similar. Simple cokriging with the linear model of corregionalization results in less smooth estimates than obtained by the other two approaches and in smaller estimation variance.

To further assess the difference, maps of the difference between collocated simple cokriging and intrinsic collocated simple cokriging and the difference between simple cokriging with the linear model of corregionalization and intrinsic collocated simple cokriging are shown on Figure 5.18. Figure 5.18 shows that collocated simple cokriging and intrinsic collocated simple cokriging give similar estimates. The estimation variances are also close for these two methods; however, intrinsic collocated simple cokriging has the same or slightly smaller estimation variance than that of collocated simple cokriging. In particular, the estimation variance obtained with collocated cokriging and intrinsic collocated cokriging are the same at the data locations (they are zero) and far from data locations (they are both equal to one); otherwise collocated simple cokriging results in higher estimation variance than intrinsic collocated simple cokriging. The slight reduction in the estimation variance in intrinsic collocated simple cokriging helps remove the variance inflation problems characteristic of collocated cokriging in sequential simulation.

From Figure 5.18 we can also note that simple cokriging with the linear model of corregionalization results in significantly different estimates than collocated cokriging and intrinsic collocated cokriging. Moreover, the estimation variance of simple cokriging in this case study for the studied covariance models is always the same or slightly smaller than the estimation variance of intrinsic collocated simple cokriging and smaller than that of collocated simple cokriging. As a result, the local conditional distributions of uncertainty obtained by the simple cokriging with the linear model of corregionalization are narrower than the local conditional distributions obtained in intrinsic collocated simple cokriging.


Difference CCK - ICCK: Mean
50.0
0.0
North
East
0.0
50.0
0.25
0.20
0.15
0.10
0.05
0.0
-0.05
-0.10
-0.15
-0.20
-0.25
Difference CCK - ICCK: Variance
0.05
0.04
0.03
0.02
0.01
0.0
Difference (CK with LMC) - ICCKM: Mean
0.75
0.60
0.45
0.30
0.15
0.0
-0.15
-0.30
-0.45
-0.60
-0.75
Difference (CK with LMC) - ICCK: Variance
0.0
-0.03
-0.06
-0.09
-0.12
-0.15
-0.18

a) b)

Figure 5.18: The maps of the difference in means (a) and variances (b) for collocated simple cokriging and simple cokriging with the intrinsic correlation model (top) and for simple cokriging with the linear model of corregionalization and simple cokriging with the intrinsic correlation model (bottom).

5.6. The Super Secondary Approach

5.6.1. Collocated Simple Cokriging with Multiple Secondary Data

Collocated simple cokriging can be implemented with multiple secondary data. The form of the estimator and estimation variance in this case under the multivariate Gaussian model are the following (e.g., Xu, et al., 1992)

$$Z^*_{CCSK}(\mathbf{u}_0) = \sum_{\alpha=1}^{n} \lambda_{Z,\alpha} Z(\mathbf{u}_\alpha) + \sum_{\mu=1}^{n_{sec}} \lambda_{Y_\mu,0} Y_\mu(\mathbf{u}_0), \tag{5.17}$$

$$\sigma^2_{CCSK}(\mathbf{u}_0) = 1 - \sum_{\alpha=1}^{n} \lambda_{Z,\alpha} \rho(\mathbf{u}_\alpha - \mathbf{u}_0) - \sum_{\mu=1}^{n_{sec}} \lambda_{Y_\mu,0} \rho_{Y_\mu,Z}, \tag{5.18}$$

where $\mathbf{u}_0$ is the estimation location, $\lambda_{Z,\alpha}$ and $\lambda_{Y_\mu,0}$ are weights applied to the primary and secondary data, respectively; $\rho(\mathbf{u}_\alpha - \mathbf{u}_0)$ are spatial correlations between primary data and the primary variable at the location being estimated; $\rho_{Y_\mu,Z}$ are correlation coefficients between primary and each secondary variable, $\alpha = 1,\ldots,n$; $\mu = 1,\ldots,n_{sec}$. The secondary data are taken at the location being estimated and the primary data are taken at other locations. The equations to compute the n+n_{sec} collocated simple cokriging weights are given below

$$\begin{aligned} &\sum_{\alpha=1}^{n} \lambda_{Z,\alpha} \rho(\mathbf{u}_\alpha - \mathbf{u}_\beta) + \sum_{\mu=1}^{n_{sec}} \lambda_{Y_\mu,0} \rho_{Y_\mu,Z}(\mathbf{u}_\beta - \mathbf{u}_0) = \rho(\mathbf{u}_\beta - \mathbf{u}_0), \quad \beta = 1,\ldots,n, \\ &\sum_{\alpha=1}^{n} \lambda_{Z,\alpha} \rho_{Y_\varsigma,Z}(\mathbf{u}_\alpha - \mathbf{u}_0) + \sum_{\mu=1}^{n_{sec}} \lambda_{Y_\varsigma,0} \rho_{Y_\varsigma,Y_\mu} = \rho_{Y_\varsigma,Z}(\mathbf{u}_\varsigma - \mathbf{u}_0), \quad \varsigma = 1,\ldots,n_{sec}, \end{aligned} \tag{5.19}$$

where the different correlation coefficients are computed from the following sources:

- $\rho(\mathbf{u}_\alpha - \mathbf{u}_\beta)$, $\alpha, \beta = 1,\ldots,n$, are spatial correlations between primary data; they are calculated directly from normal scores variogram $\gamma(\mathbf{h})$;
- $\rho_{Y_\mu,Z}(\mathbf{u}_\beta - \mathbf{u}_0)$, $\mu = 1,\ldots,n_{sec,}$ $\beta = 1,\ldots,n$, are correlations between primary and secondary data; they are calculated from the Markov model

$$\rho_{Y_\mu,Z}(\mathbf{u}_\beta - \mathbf{u}_0) = \rho_{Y_\mu,Z} \cdot \rho(\mathbf{u}_\beta - \mathbf{u}_0).$$

- ρ_{Y_ς,Y_μ}, $\mu,\varsigma = 1,\dots,n_{\text{sec},}$ are correlations between secondary data; they are calculated directly from the secondary data.

In order to perform the collocated cokriging we only require the variogram of the primary variable, which is used to compute the correlation between the primary data at other locations to the location being estimated. The spatial correlation structure of the secondary data is not required because the secondary data are only used at the location being estimated. The cross spatial correlation between primary data at other locations and the secondary data at the location being estimated are estimated through a Markov-type assumption, that is, the cross variograms are assumed to have the same shape – the sill is scaled to the correct cross correlation (Goovaerts, 1997).

5.6.2. Merging Multiple Secondary Variables (the *Super Secondary* Variable)

All secondary data can be merged as a linear combination into a single secondary variable that can be used in the conventional collocated cokriging as follows

$$Y_{\substack{\text{super}\\\text{secondary}}} = \frac{\sum_{\mu=1}^{n_{\text{sec}}} c_\mu Y_\mu}{\rho_{\substack{\text{super}\\\text{secondary}}}}, \tag{5.20}$$

where the weights c_μ are calculated from the well known normal equation of multiple linear regression:

$$\sum_{\varsigma=1}^{n_{\text{sec}}} c_\varsigma \rho_{Y_\varsigma,Y_\mu} = \rho_{Y_\mu,Z}, \quad \mu = 1,\dots,n_{\text{sec}}. \tag{5.21}$$

The left hand side correlations ρ_{Y_ς,Y_μ}, $\varsigma,\mu = 1,\dots,n_{\text{sec}}$, represent the redundancy between the secondary data; the right hand side correlations $\rho_{Y_\mu,Z}$, $\mu = 1,\dots,n_{\text{sec}}$, represent the relationship between each secondary data and the primary variable being predicted. The correlation coefficient of the super secondary variable with the primary variable being estimated is based on the cokriging variance:

$$\rho_{\substack{\text{super}\\ \text{secondary}}} = \sqrt{\sum_{\mu=1}^{n_{sec}} c_\mu \rho_{Y_\mu,Z}}\,. \tag{5.22}$$

The expression inside the square root is one minus the estimation variance, which is precisely the correlation coefficient, when only one data is being used.

The single super secondary variable is used with the primary data in the well known collocated cokriging equations:

$$Z^*_{CCSK}(\mathbf{u}_0) = \sum_{\alpha=1}^{n} \lambda_{Z,\alpha} Z(\mathbf{u}_\alpha) + \xi \cdot Y_{\substack{\text{super}\\ \text{secondary}}}(\mathbf{u}_0), \tag{5.23}$$

$$\sigma^2_{CCSK}(\mathbf{u}_0) = 1 - \sum_{\alpha=1}^{n} \lambda_{Z,\alpha} \rho(\mathbf{u}_\alpha - \mathbf{u}_0) - \xi \cdot \rho_{\substack{\text{super}\\ \text{secondary}}}\,. \tag{5.24}$$

The results of equations (5.23)-(5.24) are exactly the same as that of equations (5.17)-(5.18).

The notion of using a linear combination of attributes was proposed by Myers (1983). However, such a notion implies the impractical solution of a cumbersome and large system of equations. Collocated cokriging with merged data as developed in this paper proposes to avoid the numerical instabilities caused by the classic matrix form of cokriging (Myers 1982) which has no industrial use at present.

5.6.3. Proof of the Super Secondary Approach

Let us now prove that the two collocated cokriging estimators presented in (5.17)-(5.18) and (5.23)-(5.24) are exactly the same. First, let us rewrite both collocated cokriging Systems into matrix format. Specifically, the System (5.17)-(5.18) for cokriging with multiple secondary data can be rewritten (using Markov model) as

$$Z^*_{CCSK}(\mathbf{u}_0) = \boldsymbol{\lambda}_Z^T \mathbf{Z} + \boldsymbol{\lambda}_{Y,0}^T \mathbf{Y}, \tag{5.25}$$

$$\sigma^2_{CCSK}(\mathbf{u}_0) = 1 - \boldsymbol{\lambda}_Z^T \mathbf{r}_Z - \boldsymbol{\lambda}_{Y,0}^T \mathbf{r}_{YZ}, \tag{5.26}$$

where $\boldsymbol{\lambda}_Z^T = (\lambda_{Z,1},\ldots,\lambda_{Z,n})^T$ and $\boldsymbol{\lambda}_{Y,0}^T = (\lambda_{Y_1,0},\ldots,\lambda_{Y_{n_{sec}},0})^T$ are the weights applied in estimation to the primary $\mathbf{Z} = (Z(\mathbf{u}_1),\ldots,Z(\mathbf{u}_n))^T$ and secondary $\mathbf{Y} = (Y_1(\mathbf{u}_0),\ldots,Y_{n_{sec}}(\mathbf{u}_0))^T$ data, respectively; $\mathbf{r}_Z$ is the vector of spatial correlations

$\rho(\mathbf{u}_\alpha - \mathbf{u}_0)$; $\mathbf{r}_{YZ}$ is the vector of correlations $\rho_{Y_\mu,Z}$ between primary and multiple collocated secondary data, $\alpha = 1,\dots,n$; $\mu = 1,\dots,n_{\text{sec}}$. The weights $\boldsymbol{\lambda}_Z^T = (\lambda_{Z,1},\dots,\lambda_{Z,n})^T$ and $\boldsymbol{\lambda}_{Y,0}^T = (\lambda_{Y_1,0},\dots,\lambda_{Y_{n_{\text{sec}}},0})^T$ are given by the following System:

$$\begin{aligned} &\mathbf{R}_Z\boldsymbol{\lambda}_Z + \mathbf{r}_Z\mathbf{r}_{YZ}^T\boldsymbol{\lambda}_{Y,0} = \mathbf{r}_Z, \\ &\mathbf{r}_{YZ}\mathbf{r}_Z^T\boldsymbol{\lambda}_Z + \mathbf{R}_Y\boldsymbol{\lambda}_{Y,0} = \mathbf{r}_{YZ}, \end{aligned} \tag{5.27}$$

where $\mathbf{R}_Z$ is n by n data-to-data covariance matrix for the primary data ($\mathbf{R}_Z = \rho(\mathbf{u}_\alpha - \mathbf{u}_\beta)$, $\alpha,\beta = 1,\dots,n$); and $\mathbf{R}_Y$ is n_{sec} by n_{sec} matrix of correlations between multiple secondary data ($\mathbf{R}_Y = \rho_{Y_\varsigma,Y_\mu}$, $\mu,\varsigma = 1,\dots,n_{\text{sec}}$).

The System (5.23)-(5.24) for collocated cokriging with one super secondary data can be rewritten (using Markov model) as

$$Z^*_{CCSK}(\mathbf{u}_0) = \tilde{\boldsymbol{\lambda}}_Z^T\mathbf{Z} + \xi \cdot Y_{\substack{\text{super}\\\text{secondary}}}(\mathbf{u}_0), \tag{5.28}$$

$$\sigma^2_{CCSK}(\mathbf{u}_0) = 1 - \tilde{\boldsymbol{\lambda}}_Z^T\mathbf{r}_Z - \xi \cdot \rho_{\substack{\text{super}\\\text{secondary}}} \tag{5.29}$$

where $\tilde{\boldsymbol{\lambda}}_Z^T = (\tilde{\lambda}_{Z,1},\dots,\tilde{\lambda}_{Z,n})^T$ and ξ are the weights applied in estimation to the primary and super secondary data, respectively; as before, $\mathbf{r}_Z$ is the vector of spatial correlations $\rho(\mathbf{u}_\alpha - \mathbf{u}_0)$; $\rho_{\substack{\text{super}\\\text{secondary}}}$ is given (see Equations (5.21)-(5.22)) by

$$\rho^2_{\substack{\text{super}\\\text{secondary}}} = \mathbf{r}_{YZ}^T\mathbf{R}_Y^{-1}\mathbf{r}_{YZ}. \tag{5.30}$$

Note that $\mathbf{R}_Y$ is the correlation matrix – positive definite, thus, invertible. The weights for the primary and super secondary data, $\tilde{\boldsymbol{\lambda}}_Z^T = (\tilde{\lambda}_{Z,1},\dots,\tilde{\lambda}_{Z,n})^T$ and ξ, respectively, are found from the following System

$$\begin{aligned} &\mathbf{R}_Z\tilde{\boldsymbol{\lambda}}_Z + \mathbf{r}_Z\rho_{\substack{\text{super}\\\text{secondary}}}\,\xi = \mathbf{r}_Z, \\ &\rho_{\substack{\text{super}\\\text{secondary}}}\,\mathbf{r}_Z^T\tilde{\boldsymbol{\lambda}}_Z + \xi = \rho_{\substack{\text{super}\\\text{secondary}}}, \end{aligned} \tag{5.31}$$

where, as before, $\mathbf{R}_Z$ is n by n data-to-data covariance matrix for the primary data ($\mathbf{R}_Z = \rho(\mathbf{u}_\alpha - \mathbf{u}_\beta)$, $\alpha,\beta = 1,\dots,n$). Using the fact that (see Equations (5.20)-(5.22)):

$$Y_{\substack{\text{super}\\\text{secondary}}}(\mathbf{u}_0) = \frac{1}{\rho_{\substack{\text{super}\\\text{secondary}}}}(\mathbf{R}_Y^{-1}\mathbf{r}_{YZ})^T\mathbf{Y}, \tag{5.32}$$

we can rewrite equation for the estimate of the collocated cokriging approach with one super secondary data as:

$$Z^*_{CCSK}(\mathbf{u}_0) = \widetilde{\boldsymbol{\lambda}}_Z^T\mathbf{Z} + \xi\cdot\frac{1}{\rho_{\substack{\text{super}\\\text{secondary}}}}(\mathbf{R}_Y^{-1}\mathbf{r}_{YZ})^T\mathbf{Y}. \tag{5.33}$$

Looking at Systems (5.25)-(5.27) and (5.28)-(5.33), we can conclude that in order to prove that they result in the same outcome; we need to show that the weights received by primary and secondary data in both Systems are the same, that is, we need to show that the following equalities hold:

1) $\lambda_{Z,\alpha} = \widetilde{\lambda}_{Z,\alpha}$, $\alpha = 1,\ldots,n$;

and

2) $\lambda_{Y_\beta,0} = \xi\cdot\dfrac{1}{\rho_{\substack{\text{super}\\\text{secondary}}}}(\mathbf{R}_Y^{-1}\mathbf{r}_{YZ})_\beta$, $\beta = 1,\ldots,n_{\text{sec}}$.

Proof of 1): Let us first consider second matrix equation of System (5.27), that is,

$$\mathbf{r}_{YZ}\mathbf{r}_Z^T\boldsymbol{\lambda}_Z + \mathbf{R}_Y\boldsymbol{\lambda}_{Y,0} = \mathbf{r}_{YZ},$$

Multiplying both sides of this matrix equation by $\mathbf{r}_{YZ}^T\mathbf{R}_Y^{-1}$, we obtain:

$$\mathbf{r}_{YZ}^T\mathbf{R}_Y^{-1}\left[\mathbf{r}_{YZ}\mathbf{r}_Z^T\boldsymbol{\lambda}_Z + \mathbf{R}_Y\boldsymbol{\lambda}_{Y,0}\right] = \mathbf{r}_{YZ}^T\mathbf{R}_Y^{-1}\mathbf{r}_{YZ},$$

or,

$$\left(\mathbf{r}_{YZ}^T\mathbf{R}_Y^{-1}\mathbf{r}_{YZ}\right)\mathbf{r}_Z^T\boldsymbol{\lambda}_Z + \mathbf{r}_{YZ}^T\left(\mathbf{R}_Y^{-1}\mathbf{R}_Y\right)\boldsymbol{\lambda}_{Y,0} = \mathbf{r}_{YZ}^T\mathbf{R}_Y^{-1}\mathbf{r}_{YZ}, \tag{5.34}$$

Using (5.30) and the fact that $\mathbf{R}_Y^{-1}\mathbf{R}_Y = \mathbf{I}$, where $\mathbf{I}$ is identity matrix of size n by n, equation (5.34) reduces to:

$$\rho^2_{\substack{\text{super}\\\text{secondary}}}\mathbf{r}_Z^T\boldsymbol{\lambda}_Z + \mathbf{r}_{YZ}^T\boldsymbol{\lambda}_{Y,0} = \rho^2_{\substack{\text{super}\\\text{secondary}}}. \tag{5.35}$$

Now let us consider second matrix equation of System (5.31), that is,

$$\rho_{\substack{\text{super}\\\text{secondary}}}\mathbf{r}_Z^T\widetilde{\boldsymbol{\lambda}}_Z + \xi = \rho_{\substack{\text{super}\\\text{secondary}}}.$$

Multiplying both sides of this equation by $\rho_{\substack{\text{super}\\\text{secondary}}}$, we obtain

$$\rho^2_{\substack{\text{super}\\ \text{secondary}}} \mathbf{r}_Z^T \tilde{\boldsymbol{\lambda}}_Z + \xi \rho_{\substack{\text{super}\\ \text{secondary}}} = \rho^2_{\substack{\text{super}\\ \text{secondary}}} . \tag{5.36}$$

Subtracting from Equation (5.35) Equation (5.36), we get

$$\rho^2_{\substack{\text{super}\\ \text{secondary}}} \mathbf{r}_Z^T \boldsymbol{\lambda}_Z + \mathbf{r}_{YZ}^T \boldsymbol{\lambda}_{Y,0} - \left[\rho^2_{\substack{\text{super}\\ \text{secondary}}} \mathbf{r}_Z^T \tilde{\boldsymbol{\lambda}}_Z + \xi \rho_{\substack{\text{super}\\ \text{secondary}}} \right] = \rho^2_{\substack{\text{super}\\ \text{secondary}}} - \rho^2_{\substack{\text{super}\\ \text{secondary}}} ,$$

or,

$$\rho^2_{\substack{\text{super}\\ \text{secondary}}} \mathbf{r}_Z^T [\boldsymbol{\lambda}_Z - \tilde{\boldsymbol{\lambda}}_Z] + \left[\mathbf{r}_{YZ}^T \boldsymbol{\lambda}_{Y,0} - \xi \rho_{\substack{\text{super}\\ \text{secondary}}} \right] = 0. \tag{5.37}$$

Recall that it follows from the first matrix equations of Systems (5.27) and (5.31) that

$$\mathbf{r}_Z \mathbf{r}_{YZ}^T \boldsymbol{\lambda}_{Y,0} = \mathbf{r}_Z - \mathbf{R}_Z \boldsymbol{\lambda}_Z , \tag{5.38}$$

and

$$\mathbf{r}_Z \rho_{\substack{\text{super}\\ \text{secondary}}} \xi = \mathbf{r}_Z - \mathbf{R}_Z \tilde{\boldsymbol{\lambda}}_Z , \tag{5.39}$$

respectively. Thus, after subtracting equation (5.38) from equation (5.39), we obtain

$$\mathbf{r}_Z \mathbf{r}_{YZ}^T \boldsymbol{\lambda}_{Y,0} - \mathbf{r}_Z \rho_{\substack{\text{super}\\ \text{secondary}}} \xi = \mathbf{r}_Z - \mathbf{R}_Z \boldsymbol{\lambda}_Z - [\mathbf{r}_Z - \mathbf{R}_Z \tilde{\boldsymbol{\lambda}}_Z],$$

or,

$$\mathbf{r}_Z \left[\mathbf{r}_{YZ}^T \boldsymbol{\lambda}_{Y,0} - \xi \rho_{\substack{\text{super}\\ \text{secondary}}} \right] = -[\mathbf{R}_Z \boldsymbol{\lambda}_Z - \mathbf{R}_Z \tilde{\boldsymbol{\lambda}}_Z], \tag{5.40}$$

Let us now return to equation (5.37). If we multiply both sides of this equation by $\mathbf{r}_Z$, we obtain

$$\mathbf{r}_Z \rho^2_{\substack{\text{super}\\ \text{secondary}}} \mathbf{r}_Z^T [\boldsymbol{\lambda}_Z - \tilde{\boldsymbol{\lambda}}_Z] + \mathbf{r}_Z \left[\mathbf{r}_{YZ}^T \boldsymbol{\lambda}_{Y,0} - \xi \rho_{\substack{\text{super}\\ \text{secondary}}} \right] = 0. \tag{5.41}$$

Due to (5.40), we can rewrite (5.41) as

$$\mathbf{r}_Z \rho^2_{\substack{\text{super}\\ \text{secondary}}} \mathbf{r}_Z^T [\boldsymbol{\lambda}_Z - \tilde{\boldsymbol{\lambda}}_Z] - [\mathbf{R}_Z \boldsymbol{\lambda}_Z - \mathbf{R}_Z \tilde{\boldsymbol{\lambda}}_Z] = 0,$$

or,

$$\left[\mathbf{r}_Z \rho^2_{\substack{\text{super}\\ \text{secondary}}} \mathbf{r}_Z^T - \mathbf{R}_Z \right] [\boldsymbol{\lambda}_Z - \tilde{\boldsymbol{\lambda}}_Z] = 0. \tag{5.42}$$

Thus, it follows from equation (5.42) that $\boldsymbol{\lambda}_Z = \widetilde{\boldsymbol{\lambda}}_Z$ provided that determinant of matrix $\mathbf{r}_Z \rho^2_{\substack{\text{super}\\\text{secondary}}} \mathbf{r}_Z^T - \mathbf{R}_Z$ is not equal to zero. So, let us examine whether determinant of matrix $\mathbf{r}_Z \rho^2_{\substack{\text{super}\\\text{secondary}}} \mathbf{r}_Z^T - \mathbf{R}_Z$ could be zero. Consider System (5.31) rewritten below,

$$\begin{aligned} &\mathbf{R}_Z \widetilde{\boldsymbol{\lambda}}_Z + \mathbf{r}_Z \rho_{\substack{\text{super}\\\text{secondary}}} \xi = \mathbf{r}_Z, \\ &\xi = \rho_{\substack{\text{super}\\\text{secondary}}} [1 - \mathbf{r}_Z^T \widetilde{\boldsymbol{\lambda}}_Z]. \end{aligned} \tag{5.43}$$

If we substitute expression for ξ from second equation of System (5.43) into the first matrix equation of this System, we will obtain the following

$$\begin{aligned} &\mathbf{R}_Z \widetilde{\boldsymbol{\lambda}}_Z + \mathbf{r}_Z \rho^2_{\substack{\text{super}\\\text{secondary}}} [1 - \mathbf{r}_Z^T \widetilde{\boldsymbol{\lambda}}_Z] = \mathbf{r}_Z, \\ &\xi = \rho_{\substack{\text{super}\\\text{secondary}}} [1 - \mathbf{r}_Z^T \widetilde{\boldsymbol{\lambda}}_Z], \end{aligned}$$

or,

$$\begin{aligned} &-\left[\mathbf{r}_Z \rho^2_{\substack{\text{super}\\\text{secondary}}} \mathbf{r}_Z^T - \mathbf{R}_Z\right] \widetilde{\boldsymbol{\lambda}}_Z = \mathbf{r}_Z \left[1 - \rho^2_{\substack{\text{super}\\\text{secondary}}}\right], \\ &\xi = \rho_{\substack{\text{super}\\\text{secondary}}} [1 - \mathbf{r}_Z^T \widetilde{\boldsymbol{\lambda}}_Z], \end{aligned} \tag{5.44}$$

Thus, clearly, if determinant of matrix $\mathbf{r}_Z \rho^2_{\substack{\text{super}\\\text{secondary}}} \mathbf{r}_Z^T - \mathbf{R}_Z$ is equal to zero, the System (5.44) will have either multiple or no solution depending on the vector $\mathbf{r}_Z \left[1 - \rho^2_{\substack{\text{super}\\\text{secondary}}}\right]$ (Cramer's rule). The collocated cokriging System with super secondary variable (5.31) (or, equivalently (5.43)) has only one solution for the primary variable weights $\widetilde{\boldsymbol{\lambda}}_Z$ if and only if determinant of matrix $\mathbf{r}_Z \rho^2_{\substack{\text{super}\\\text{secondary}}} \mathbf{r}_Z^T - \mathbf{R}_Z$ is not equal to zero. Similarly, using simple matrix manipulations, we can show that the collocated cokriging System for the weights (5.27) will have unique solution for the primary weights if and only if determinant of matrix $\mathbf{r}_Z \rho^2_{\substack{\text{super}\\\text{secondary}}} \mathbf{r}_Z^T - \mathbf{R}_Z$ is not equal to zero. As a result, we can conclude that provided that the collocated cokriging Systems each have unique solution, the weights given by both collocated cokriging Systems (5.25)-(5.27) and (5.28)-(5.33) to the primary variable are exactly the same. That is,

$$\lambda_{Z,\alpha} = \tilde{\lambda}_{Z,\alpha}, \quad \alpha = 1,\ldots,n.$$

Thus, proof of 1) is completed.

Proof of 2): Let us now prove that

$$\lambda_{Y_\beta,0} = \xi \cdot \frac{1}{\rho_{\substack{\text{super}\\ \text{secondary}}}} (\mathbf{R}_Y^{-1}\mathbf{r}_{YZ})_\beta, \quad \beta = 1,\ldots,n_{\text{sec}}.$$

It follows from the second matrix equation of System (5.27) for the collocated cokriging weights with multiple secondary data that

$$\mathbf{R}_Y \boldsymbol{\lambda}_{Y,0} = \mathbf{r}_{YZ} - \mathbf{r}_{YZ}\mathbf{r}_Z^T\boldsymbol{\lambda}_Z = \mathbf{r}_{YZ}[1 - \mathbf{r}_Z^T\boldsymbol{\lambda}_Z].$$

Also, it follows from System (5.44) (rewritten System (5.31)) for the collocated cokriging weights with super secondary data that

$$\xi = \rho_{\substack{\text{super}\\ \text{secondary}}} [1 - \mathbf{r}_Z^T\tilde{\boldsymbol{\lambda}}_Z].$$

As result,

$$\begin{aligned}
\mathbf{R}_Y\left[\boldsymbol{\lambda}_{Y,0} - \xi \cdot \frac{1}{\rho_{\substack{\text{super}\\ \text{secondary}}}}(\mathbf{R}_Y^{-1}\mathbf{r}_{YZ})\right] &= \mathbf{R}_Y\boldsymbol{\lambda}_{Y,0} - \mathbf{R}_Y\xi \cdot \frac{1}{\rho_{\substack{\text{super}\\ \text{secondary}}}}(\mathbf{R}_Y^{-1}\mathbf{r}_{YZ}) \\
&= \mathbf{R}_Y\boldsymbol{\lambda}_{Y,0} - \xi \cdot \frac{1}{\rho_{\substack{\text{super}\\ \text{secondary}}}}(\mathbf{R}_Y\mathbf{R}_Y^{-1})\mathbf{r}_{YZ} = \mathbf{R}_Y\boldsymbol{\lambda}_{Y,0} - \xi \cdot \frac{1}{\rho_{\substack{\text{super}\\ \text{secondary}}}}\mathbf{r}_{YZ} \\
&= \mathbf{r}_{YZ}[1 - \mathbf{r}_Z^T\boldsymbol{\lambda}_Z] - \rho_{\substack{\text{super}\\ \text{secondary}}}[1 - \mathbf{r}_Z^T\tilde{\boldsymbol{\lambda}}_Z]\frac{1}{\rho_{\substack{\text{super}\\ \text{secondary}}}}\mathbf{r}_{YZ} \\
&= \mathbf{r}_{YZ}[1 - \mathbf{r}_Z^T\boldsymbol{\lambda}_Z] - [1 - \mathbf{r}_Z^T\tilde{\boldsymbol{\lambda}}_Z]\mathbf{r}_{YZ} = [1 - \mathbf{r}_Z^T\tilde{\boldsymbol{\lambda}}_Z]\mathbf{r}_{YZ} - [1 - \mathbf{r}_Z^T\tilde{\boldsymbol{\lambda}}_Z]\mathbf{r}_{YZ} = 0.
\end{aligned} \tag{5.45}$$

Since matrix $\mathbf{R}_Y$ is a positive-definite matrix of correlations between collocated secondary data, its determinant is not equal to zero. As result, it follows from (5.45) that $\boldsymbol{\lambda}_{Y,0} - \xi \cdot \frac{1}{\rho_{\substack{\text{super}\\ \text{secondary}}}}(\mathbf{R}_Y^{-1}\mathbf{r}_{YZ}) = 0$. Therefore, we have proved that

$$\lambda_{Y_\beta,0} = \xi \cdot \frac{1}{\rho_{\substack{\text{super}\\ \text{secondary}}}}(\mathbf{R}_Y^{-1}\mathbf{r}_{YZ})_\beta, \quad \beta = 1,\ldots,n_{\text{sec}}.$$

And, thus, we have shown that the two collocated cokriging estimators and their variances presented in (5.25)-(5.27) and (5.28)-(5.33) are exactly the same.

5.7. Extension of the Super Secondary Approach to Intrinsic Collocated Cokriging

5.7.1. Intrinsic Collocated Cokriging with Multiple Secondary Data

Intrinsic collocated simple cokriging can be implemented with multiple secondary data. The intrinsic collocated cokriging estimator and its estimation variance under the multivariate Gaussian model are given below

$$Z^*_{ICCK}(\mathbf{u}_0) = \sum_{\alpha=1}^{N} \lambda_{Z,\alpha} Z(\mathbf{u}_\alpha) + \sum_{\mu=1}^{n_{sec}} \lambda_{Y_\mu,0} Y_\mu(\mathbf{u}_0) + \sum_{\mu=1}^{n_{sec}} \sum_{\alpha=1}^{N} \lambda_{Y_\mu,\alpha} Y_\mu(\mathbf{u}_\alpha), \qquad (5.50)$$

$$\sigma^2_{ICCK}(\mathbf{u}_0) = 1 - \sum_{\alpha=1}^{n} \lambda_{Z,\alpha} \rho(\mathbf{u}_\alpha - \mathbf{u}_0) - \sum_{\mu=1}^{n_{sec}} \lambda_{Y_\mu,0} \rho_{Y_\mu,Z} - \sum_{\mu=1}^{n_{sec}} \sum_{\alpha=1}^{N} \lambda_{Y_\mu,\alpha} \rho_{Y_\mu,Z}(\mathbf{u}_\alpha - \mathbf{u}_0), \qquad (5.51)$$

where $\mathbf{u}_0$ is the estimation location, $\lambda_{Z,\alpha}$, $\lambda_{Y_\mu,\alpha}$ and $\lambda_{Y_\mu,0}$ are weights applied to the primary, secondary and collocated secondary data, respectively; $\rho(\mathbf{u}_\alpha - \mathbf{u}_0)$ are spatial correlations between primary data and the primary variable at the location being estimated; $\rho_{Y_\mu,Z}$ are correlation coefficients between primary and each collocated secondary variable; $\rho_{Y_\mu,Z}(\mathbf{u}_\beta - \mathbf{u}_0) = \rho_{Y_\mu,Z}\rho(\mathbf{u}_\beta - \mathbf{u}_0)$ are spatial correlations between secondary data and primary data at the estimation location, $\alpha = 1,\ldots,n$; $\mu = 1,\ldots,n_{sec}$.

The equations to compute the $n + n_{sec}(n+1)$ intrinsic collocated simple cokriging weights are given below

$$\begin{aligned}
&\sum_{\alpha=1}^{n} \lambda_{Z,\alpha} \rho(\mathbf{u}_\alpha - \mathbf{u}_\beta) + \sum_{\mu=1}^{n_{sec}} \sum_{\alpha=1}^{n} \lambda_{Y\mu,\alpha} \rho_{Y_\mu,Z}(\mathbf{u}_\alpha - \mathbf{u}_\beta) + \sum_{\mu=1}^{n_{sec}} \lambda_{Y_\mu,0} \rho_{Y_\mu,Z}(\mathbf{u}_\beta - \mathbf{u}_0) = \rho(\mathbf{u}_\beta - \mathbf{u}_0), \quad \beta = 1,\ldots,n, \\
&\sum_{\alpha=1}^{n} \lambda_{Z,\alpha} \rho_{Y_\varsigma,Z}(\mathbf{u}_\alpha - \mathbf{u}_\beta) + \sum_{\mu=1}^{n_{sec}} \sum_{\alpha=1}^{n} \lambda_{Y_\mu,\alpha} \rho_{Y_\mu,Y_\varsigma}(\mathbf{u}_\alpha - \mathbf{u}_\beta) + \sum_{\mu=1}^{n_{sec}} \lambda_{Y_\mu,0} \rho_{Y_\mu,Y_\varsigma}(\mathbf{u}_\beta - \mathbf{u}_0) = \rho_{Y_\varsigma,Z}(\mathbf{u}_\beta - \mathbf{u}_0), \\
&\hspace{20em} \beta = 1,\ldots,n, \quad \varsigma = 1,\ldots,n_{sec}, \\
&\sum_{\alpha=1}^{n} \lambda_{Z,\alpha} \rho_{Y_\varsigma,Z}(\mathbf{u}_\alpha - \mathbf{u}_0) + \sum_{\mu=1}^{n_{sec}} \sum_{\alpha=1}^{n} \lambda_{Y_\mu,\alpha} \rho_{Y_\mu,Y_\varsigma}(\mathbf{u}_\alpha - \mathbf{u}_0) + \sum_{\mu=1}^{n_{sec}} \lambda_{Y_\varsigma,0} \rho_{Y_\varsigma,Y_\mu} = \rho_{Y_\varsigma,Z}(\mathbf{u}_\varsigma - \mathbf{u}_0), \quad \varsigma = 1,\ldots,n_{sec},
\end{aligned} \qquad (5.52)$$

where the different correlation coefficients are calculated as

- $\rho(\mathbf{u}_\alpha - \mathbf{u}_\beta)$, $\alpha, \beta = 1,\ldots,n$, are spatial correlations between primary data; they are calculated directly from normal scores variogram $\gamma(\mathbf{h})$;
- $\rho_{Y_\mu,Z}(\mathbf{u}_\beta - \mathbf{u}_0) = \rho_{Y_\mu,Z}\rho(\mathbf{u}_\beta - \mathbf{u}_0)$ and $\rho_{Y_\mu,Z}(\mathbf{u}_\beta - \mathbf{u}_\alpha) = \rho_{Y_\mu,Z}\rho(\mathbf{u}_\beta - \mathbf{u}_\alpha)$, $\mu = 1,\ldots,n_{\text{sec}}$, $\alpha, \beta = 1,\ldots,n$, are spatial correlations between primary and secondary data.
- $\rho_{Y_\varsigma,Y_\mu}(\mathbf{u}_\alpha - \mathbf{u}_\beta) = \rho_{Y_\varsigma,Y_\mu}\rho(\mathbf{u}_\alpha - \mathbf{u}_\beta)$, $\mu, \varsigma = 1,\ldots,n_{\text{sec}}$, $\alpha, \beta = 0,\ldots,n$, are correlations between secondary data.

In order to perform the intrinsic collocated cokriging, we only require the variogram of the primary variable and the matrix of the correlation coefficients between multiple secondary data and primary and secondary data.

5.7.2. Intrinsic Collocated Cokriging with Super Secondary Variable

All secondary data can be merged as a linear combination into a single secondary variable using Equations (5.20)-(5.22). Then this super secondary variable can be used with the primary data in the single variable intrinsic collocated cokriging System:

$$Z^*_{ICCK}(\mathbf{u}_0) = \sum_{\alpha=1}^{N} \lambda_{Z,\alpha} Z(\mathbf{u}_\alpha) + \lambda_{Y_{\substack{\text{super}\\\text{secondary}}},0} Y_{\substack{\text{super}\\\text{secondary}}}(\mathbf{u}_0) + \sum_{\alpha=1}^{N} \lambda_{Y_{\substack{\text{super}\\\text{secondary}}},\alpha} Y_{\substack{\text{super}\\\text{secondary}}}(\mathbf{u}_\alpha), \quad (5.53)$$

$$\sigma^2_{ICCK}(\mathbf{u}_0) = 1 - \sum_{\alpha=1}^{n} \lambda_{Z,\alpha} \rho(\mathbf{u}_\alpha - \mathbf{u}_0) - \lambda_{Y_{\substack{\text{super}\\\text{secondary}}},0} \rho_{\substack{\text{super}\\\text{secondary}}} - \sum_{\alpha=1}^{N} \lambda_{Y_{\substack{\text{super}\\\text{secondary}}},\alpha} \rho_{Y_{\substack{\text{super}\\\text{secondary}}},Z}(\mathbf{u}_\alpha - \mathbf{u}_0), \quad (5.54)$$

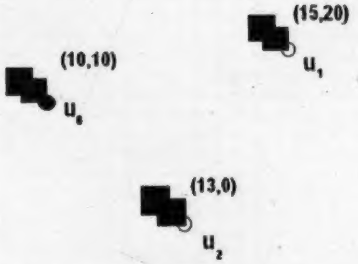
where

$$\rho_{Y_{\substack{\text{super}\\\text{secondary}}},Z}(\mathbf{u}_\alpha - \mathbf{u}_0) = \rho_{\substack{\text{super}\\\text{secondary}}} \rho(\mathbf{u}_\alpha - \mathbf{u}_0) \quad (5.55)$$

The results of equations (5.50)-(5.51) are exactly the same as that of equations (5.53)-(5.54). This will be verified with a small example. A full proof would follow the logic of Section 5.6.1.

5.7.3. Example

Let us consider the following data configuration:



where at data locations $\mathbf{u}_1$ and $\mathbf{u}_2$ both primary (yellow circles) and secondary information of two different types (red and blue squares) are available, while at the estimation location $\mathbf{u}_0$ only secondary data values of two different types are given. Precise information given in these three locations is summarized below. The data are in Gaussian units.

	$\mathbf{u}_1$	$\mathbf{u}_2$	$\mathbf{u}_0$
Secondary I (Y_1)	-2.5	-0.5	-2
Secondary II (Y_2)	-1.5	-0.25	-0.5
Primary (Z)	-2	-1	*unknown*

Also assume that the primary variable variogram is isotropic spherical with range of continuity equal to 10, the correlation between primary and each secondary variable and between secondary variables are given below.

$$\rho_{Z,Y_1} = 0.6; \quad \rho_{Z,Y_2} = 0.4; \quad \rho_{Y_1,Y_2} = 0.8.$$

Let us now estimate the value of the primary variable at the estimation location $\mathbf{u}_0$ based on all available information using full intrinsic collocated cokriging (two secondary variables are not merged) and intrinsic collocated cokriging with one super secondary variable and verify that results are the same.

Full intrinsic collocated cokriging primary variable estimator and its estimation variance at location $\mathbf{u}_0$ is given by (see Equations (5.50)-(5.51))

$$Z^*_{ICCK}(\mathbf{u}_0) = \sum_{\alpha=1}^{2} \lambda_{Z,\alpha} Z(\mathbf{u}_\alpha) + \sum_{\mu=1}^{2} \lambda_{Y_\mu,0} Y_\mu(\mathbf{u}_0) + \sum_{\mu=1}^{2} \sum_{\alpha=1}^{2} \lambda_{Y_\mu,\alpha} Y_\mu(\mathbf{u}_\alpha) \qquad (5.56)$$

$$\sigma^2_{ICCK}(\mathbf{u}_0) = 1 - \sum_{\alpha=1}^{2} \lambda_{Z,\alpha} \rho(\mathbf{u}_\alpha - \mathbf{u}_0) - \sum_{\mu=1}^{2} \lambda_{Y_\mu,0} \rho_{Y_\mu,Z} - \sum_{\mu=1}^{2} \sum_{\alpha=1}^{2} \lambda_{Y_\mu,\alpha} \rho_{Y_\mu,Z}(\mathbf{u}_\alpha - \mathbf{u}_0) \quad (5.57)$$

The weights for the above System can be calculated from System (5.52) with $n = 2$ and $n_{\text{sec}} = 2$.

After solving System (5.52) and substituting resulting weights into equations (5.56)-(5.57), we obtain the following full intrinsic collocated cokriging primary variable estimate and estimation variance at location $\mathbf{u}_0$

$$Z^*_{ICCK}(\mathbf{u}_0) = -1.444 \qquad \text{and} \qquad \sigma^2_{ICCK}(\mathbf{u}_0) = 0.622. \quad (5.58)$$

Intrinsic collocated cokriging with super secondary variable. The super secondary variable for intrinsic collocated cokriging can be calculated as follows (see Equation (5.20))

$$Y_{\substack{\text{super}\\\text{secondary}}} = \frac{\sum_{\mu=1}^{2} c_\mu Y_\mu}{\rho_{\substack{\text{super}\\\text{secondary}}}}, \quad (5.59)$$

where the weights calculated from System (5.21) are given by

$$c_1 = 0.778 \quad \text{and} \quad c_2 = -0.222; \quad (5.60)$$

and

$$\rho_{\substack{\text{super}\\\text{secondary}}} = \sqrt{\sum_{\mu=1}^{2} c_\mu \rho_{Y_\mu,Z}} = \sqrt{0.778 \cdot 0.6 - 0.222 \cdot 0.4} \doteq 0.615. \quad (5.61)$$

Thus, the super secondary variable is given by

$$Y_{\substack{\text{super}\\\text{secondary}}} = 1.265\, Y_1 - 0.362\, Y_2, \quad (5.62)$$

and the intrinsic collocated cokriging estimator of primary variable and its estimation variance at location $\mathbf{u}_0$ based on the super secondary approach is given by (see equations (5.53)-(5.54))

$$Z^*_{ICCK}(\mathbf{u}_0) = \sum_{\alpha=1}^{N} \lambda_{Z,\alpha} Z(\mathbf{u}_\alpha) + \lambda_{Y_{\substack{\text{super}\\\text{secondary}}},0} Y_{\substack{\text{super}\\\text{secondary}}}(\mathbf{u}_0) + \sum_{\alpha=1}^{N} \lambda_{Y_{\substack{\text{super}\\\text{secondary}}},\alpha} Y_{\substack{\text{super}\\\text{secondary}}}(\mathbf{u}_\alpha), \quad (5.63)$$

$$\sigma^2_{ICCK}(\mathbf{u}_0) = 1 - \sum_{\alpha=1}^{n} \lambda_{Z,\alpha} \rho(\mathbf{u}_\alpha - \mathbf{u}_0) - \lambda_{Y_{\substack{\text{super}\\\text{secondary}}},0} \rho_{\substack{\text{super}\\\text{secondary}}} - \sum_{\alpha=1}^{N} \lambda_{Y_{\substack{\text{super}\\\text{secondary}}},\alpha} \rho_{Y_{\substack{\text{super}\\\text{secondary}}},Z}(\mathbf{u}_\alpha - \mathbf{u}_0). \quad (5.64)$$

After calculating the intrinsic collocated cokriging weights and substituting them into equations (5.53)-(5.54), we obtain the following intrinsic collocated cokriging estimator and its estimation variance at location $\mathbf{u}_0$

$$Z^*_{ICCK}(\mathbf{u}_0) = -1.444 \qquad \text{and} \qquad \sigma^2_{ICCK}(\mathbf{u}_0) = 0.622. \tag{5.65}$$

The results of the full intrinsic collocated cokriging (5.58) and intrinsic collocated cokriging with one super secondary variable (5.65) are exactly the same.

5.8. Discussion

Collocated cokriging is widely used in simulation because of its simplicity. The original Markov models with a single secondary data are prone to variance inflation leading to potential biases in the predictions. An intrinsic model of coregionalization and the use of secondary data at all primary data locations is proposed in this chapter to deal with this problem of variance inflation. The proposed approach employs full simple cokriging based on the intrinsic coregionalization model to calculate local distributions. Theoretical results and small examples demonstrate that the new methodology removes variance inflation, insures reproduction of the correlation between primary and secondary data and improves the reproduction of the variogram even when the primary and secondary variables differ significantly in continuity.

Moreover, another interesting development of this chapter is a super secondary approach. Geostatistical modeling is improved when estimation is constrained to all available secondary data. Within the proposed super secondary framework the multiple secondary data are merged into a *merged* secondary variable. Then, intrinsic collocated cokriging can be used with the one merged variable. Under the Markov coregionalization model, this is exactly equivalent to multivariate intrinsic collocated cokriging.

CHAPTER 6

Multiple Univariate SGS Honoring a Correlation Matrix

This chapter investigates multiple univariate sequential Gaussian simulation. A new approach for correcting the multiple univariate sequential Gaussian simulation to honor the correlation between random variables at lag 0 is proposed. The correction approach is shown to perform well in several small examples and a case study.

Section 6.1 reviews options for multivariate sequential Gaussian simulation. **Section 6.2** presents a correction to the multiple univariate sequential Gaussian simulation to honor the correlation between random variables at lag 0. **Section 6.3** and **Section 6.4** show application of the developed correction technique in several small examples and a case study. **Section 6.5** presents extension of the multiple univariate sequential Gaussian simulation to honor locally correlation between random variables at lag distance 0. A brief discussion of the results and a correction method is presented in **Section 6.6**.

6.1 Multivariate Sequential Gaussian Simulation

Univariate sequential Gaussian simulation can be easily extended to simultaneous modeling of several random variables. Then there are several options to consider.

Firstly, if a joint model of the spatial continuity for the N variables is available, then sequential Gaussian simulation with simple cokriging can be applied to multivariate simulation (see Chapter 2). Another possibility would be to apply sequential simulation with either collocated cokriging or intrinsic collocated cokriging (see Chapter 5 for details). However, if the secondary data information is not exhaustively sampled or the goal of simulation is generating multiple correlated primary variables there is also another option equally simple as SGS with intrinsic collocated cokriging. It is based on independent modeling of each random variable, but sampling with correlated residuals. The data on each individual random variable is used to calculate the mean and variance of the local conditional distribution for that variable via simple kriging. Simple kriging is performed as many times as there are variables. The variogram models to be used in multiple univariate simulation are modeled independently from each other; no inference of the joint model for spatial continuity is needed. Note, however, that this approach is a practical shortcut, and should not be considered as a replacement of full simple cokriging with a linear model of coregionalizetion.

Before we explain multiple univariate SGS with correlated residuals, let us consider independent multiple univariate sequential Gaussian simulation (that is, multiple univariate SGS with independent residuals) first. The following algorithm can be devised for independent multiple univariate sequential Gaussian simulation:

1. Transform the N variables into the standard Normal.
2. Select a random path to visit each location to be simulated.
3. Visit each location one by one and perform simple kriging for each variable to find the mean $m_{SK,i}(\mathbf{u})$ and variance $\sigma^2_{SK,i}(\mathbf{u})$ of the local conditional distributions for all N variables.
4. Draw a vector of independent standard normal random values $\mathbf{R}(\mathbf{u})^T = [R_1(\mathbf{u}),\ldots,R_N(\mathbf{u})]^T$. Then, calculate the vector of simulated values $\mathbf{Z}(\mathbf{u})^T = [Z_1(\mathbf{u}),\ldots,Z_N(\mathbf{u})]^T$ for each of the N random variables at the unsampled location $\mathbf{u}$ as

$$\mathbf{Z}(\mathbf{u}) = \mathbf{m}_{SK}(\mathbf{u}) + \boldsymbol{\sigma}_{SK}(\mathbf{u})\mathbf{R}(\mathbf{u}), \tag{6.1}$$

where $\mathbf{m}_{SK}(\mathbf{u})^T = [m_{SK,1}(\mathbf{u}),\ldots,m_{SK,N}(\mathbf{u})]^T$ is the vector of simple kriging means; matrix $\boldsymbol{\sigma}_{SK}(\mathbf{u})$ is given by

$$\boldsymbol{\sigma}_{SK}(\mathbf{u}) = \begin{bmatrix} \sigma_{SK,1}(\mathbf{u}) & 0 & \cdots & 0 \\ 0 & \sigma_{SK,2}(\mathbf{u}) & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \sigma_{SK,N}(\mathbf{u}) \end{bmatrix}. \qquad (6.2)$$

where values on the diagonal are equal to the square root of the simple kriging variances $\sigma^2_{SK,i}(\mathbf{u})$ obtained for the N variables in step 3.

5. Add the simulated values to the database to be used in simulation of subsequent nodes.
6. Repeat steps 3-5 until all locations are populated.

Multiple equally-probable realizations can be created by changing the random number seed, that is, changing a random path and selecting different set of realizations. The shape of the conditional distributions is Gaussian ensuring that the simulated realizations for all variables will be standard normally distributed. The variogram reproduction is ensured by using not only data values in simple kriging but also simulated nodes.

The correlation between the different random variables can be introduced into multiple univariate SGS by generating correlated residuals $\mathbf{R}(\mathbf{u})^T = [R_1(\mathbf{u}),\ldots,R_N(\mathbf{u})]^T$. The residuals $\mathbf{R}(\mathbf{u})^T = [R_1(\mathbf{u}),\ldots,R_N(\mathbf{u})]^T$ will be independent from location to location, that is,

$$\rho(R_i(\mathbf{u}_k), R_j(\mathbf{u}_l)) = Cov(R_i(\mathbf{u}_k), R_j(\mathbf{u}_l)) = 0, \quad \forall i,j = 1,\ldots,N, \quad \forall k \neq l; \qquad (6.3)$$

but correlated at the same location with correlation matrix $\boldsymbol{\rho}$, that is,

$$\rho(R_i(\mathbf{u}_k), R_j(\mathbf{u}_k)) = Cov(R_i(\mathbf{u}_k), R_j(\mathbf{u}_k)) = \rho_{ij}, \quad \forall i,j = 1,\ldots,N, \quad \forall k. \qquad (6.4)$$

6.2. Reproducing the Correlation Matrix between Variables at Lag 0

The global correlation matrix for the residuals, $\boldsymbol{\rho}$, to be used in multiple univariate SGS can be found based on the target correlation matrix for the random variables to be simulated, $\boldsymbol{\rho}^{\text{target}}$. The procedure is as follows.

Let us first calculate the correlation between two simulated values for variables Z_i and Z_j, respectively, at arbitrary location $\mathbf{u}$ in the study domain, $i, j \in \{1,\ldots,N\}$. These simulated values $Z_i(\mathbf{u})$ and $Z_j(\mathbf{u})$ in sequential Gaussian simulation are given by

$$\begin{aligned} Z_i(\mathbf{u}) &= m_{SK,i}(\mathbf{u}) + \sigma_{SK,i}(\mathbf{u})R_i(\mathbf{u}), \\ Z_j(\mathbf{u}) &= m_{SK,j}(\mathbf{u}) + \sigma_{SK,j}(\mathbf{u})R_j(\mathbf{u}), \end{aligned} \tag{6.5}$$

then due to independence between vector of residuals $[R_i(\mathbf{u})\ R_j(\mathbf{u})]^T$ and vector of simple kriging means $[m_{SK,i}(\mathbf{u})\ m_{SK,j}(\mathbf{u})]^T$,

$$\begin{aligned} \rho(Z_i(\mathbf{u}), Z_j(\mathbf{u})) &= Cov(Z_i(\mathbf{u}), Z_j(\mathbf{u})) \\ &= Cov(m_{SK,i}(\mathbf{u}), m_{SK,j}(\mathbf{u})) + \sigma_{SK,i}(\mathbf{u})\sigma_{SK,j}(\mathbf{u})Cov(R_i(\mathbf{u}), R_j(\mathbf{u})) \\ &= Cov(m_{SK,i}(\mathbf{u}), m_{SK,j}(\mathbf{u})) + \sigma_{SK,i}(\mathbf{u})\sigma_{SK,j}(\mathbf{u})\rho_{ij}. \end{aligned} \tag{6.6}$$

Therefore, it follows from (6.6) that the average correlation between variables Z_i and Z_j is equal to

$$\overline{\rho(Z_i(\mathbf{u}), Z_j(\mathbf{u}))} = \overline{Cov(m_{SK,i}(\mathbf{u}), m_{SK,j}(\mathbf{u}))} + \overline{\sigma_{SK,i}(\mathbf{u})\sigma_{SK,j}(\mathbf{u})}\rho_{ij};$$

the average is taken over all locations $\mathbf{u}$ in the study domain.

Then in order for the multiple univariate sequential Gaussian simulation to reproduce the target correlation between variables Z_i and Z_j at lag 0, the following equality must hold

$$\rho_{ij}^{\text{target}} = \overline{\rho(Z_i(\mathbf{u}), Z_j(\mathbf{u}))} = \overline{Cov(m_{SK,i}(\mathbf{u}), m_{SK,j}(\mathbf{u}))} + \overline{\sigma_{SK,i}(\mathbf{u})\sigma_{SK,j}(\mathbf{u})}\rho_{ij}. \tag{6.7}$$

Because the simple kriging mean values, $m_{SK,i}(\mathbf{u})$ and $m_{SK,j}(\mathbf{u})$ at any location of the study domain $\mathbf{u}$ are calculated based on the conditioning data and previously simulated nodes, the covariance between variables Z_i and Z_j at any location $\mathbf{u}$ within study domain is a linear function in the correlation coefficient ρ_{ij}. Thus, we can rewrite Equation (6.7) for the global correlation between random variables Z_i and Z_j as follows

$$\rho_{ij}^{\text{target}} = a_{ij} + b_{ij}\rho_{ij}, \tag{6.8}$$

where a_{ij} and b_{ij}, $i,j = 1,...,N$ are constants calculated based on the Simple Kriging weights (and variances). Note that if the multiple univariate sequential Gaussian simulation is unconditional, then constant a_{ij} in Equation (6.8) is equal to 0 and

$$\rho_{ij}^{\text{target}} = b_{ij}\rho_{ij}. \tag{6.9}$$

The simplest approach for finding the constants a_{ij} and b_{ij}, $i,j = 1,...,N$, in Equation (6.8) is the following:

1. Independent ($\rho_{ij} = 0, \quad i \neq j = 1,\ldots,N$) multiple univariate SGS realizations are generated to find a_{ij}'s as the correlations between realizations for variables Z_i, Z_j, $i,j = 1,...,N$. Due to ergodic fluctuations, we expect a minor change in the result for correlations between realizations for different random variables. Therefore the value for each a_{ij}, $(i \neq j) = 1,...,N$, is taken as an average correlation between realizations for variables Z_i and Z_j.
2. Fully dependent ($R_i(\mathbf{u}) = r, \quad i = 1,\ldots,N$, where r is a standard normal residual value generated for location $\mathbf{u}$) multiple univariate SGS is generated to find b_{ij}'s as the average (over all realizations) difference between the correlation of random variables Z_i and Z_j and a_{ij}'s from (1), $(i \neq j) = 1,...,N$.

Thus, in order for the multiple univariate sequential Gaussian simulation to honor the correlation matrix between random variables at lag 0, $\boldsymbol{\rho}^{\text{target}}$, the residuals with the following correlation structure need to be generated at each step of the simulation

$$\begin{aligned} &\rho(R_i(\mathbf{u}), R_j(\mathbf{u})) = \rho_{ij} = \begin{cases} \dfrac{\rho_{ij}^{\text{target}} - a_{ij}}{b_{ij}}, & \text{if} \left| \dfrac{\rho_{ij}^{\text{target}} - a_{ij}}{b_{ij}} \right| \leq 1, \quad \forall (i \neq j) = 1,...,N; \\ 1 \ \text{ or } \ -1, & \text{otherwise}; \end{cases} \\ &\rho(R_i(\mathbf{u}), R_i(\mathbf{u})) = \rho_{ii} = 1, \quad \forall i = 1,...,N. \end{aligned} \tag{6.10}$$

It is important to note that the correction in Equation (6.10) results only in global reproduction of the target collocated correlation matrix, $\boldsymbol{\rho}^{\text{target}}$; this correlation matrix may not be reproduced locally due to correlation between variables Z_i and Z_j at lag 0 being non-stationary (see Equation (6.6)).

An unfortunate feature of the correction (6.10) is that the combined matrix of the correlations $\boldsymbol{\rho}$ to be used in the multiple univariate SGS to reproduce the target correlation $\boldsymbol{\rho}^{\text{target}}$ is not necessarily positive definite. This is connected to the fact that the correlation between some variables needs to be magnified more (b_{ij} is smaller) than between others. Therefore, if matrix (6.10) is not positive definite, a positive definiteness correction to this matrix must be applied.

6.3. Unconditional Multiple Univariate SGS: Examples

6.3.1. Example 1

Let us consider the two standard normal random variables Z_1 and Z_2 with the following variograms characterizing their spatial continuity

$$\begin{aligned}\gamma_{Z_1}(\mathbf{h}) &= 0.1\cdot Sph_{16}(\mathbf{h}) + 0.9\cdot Gaus_{32}(\mathbf{h})\\ \gamma_{Z_2}(\mathbf{h}) &= 0.9\cdot Sph_{16}(\mathbf{h}) + 0.1\cdot Gaus_{32}(\mathbf{h})\end{aligned}, \tag{6.11}$$

The correlation between the two random variables is assumed to be 0.5. Note the two random variables under study have very different spatial continuity. This example was specifically chosen to show that multiple univariate SGS can be efficiently used to simulate random variables with different continuity structures while reproducing even reasonably high correlation between them. The conclusions drawn from this example are considered general.

Let us now apply multiple univariate sequential Gaussian simulation to generate 100 unconditional realizations of the two random variables on the grid of 256 by 256 blocks of size 1 by 1 unit using their respective variogram models, that is, $\gamma_{Z_1}(\mathbf{h})$ and $\gamma_{Z_2}(\mathbf{h})$, so that the correlation at lag 0 between Z_1 and Z_2 is 0.5.

To find the correlation coefficient ρ_{12} to be used in multiple univariate SGS to reproduce the target correlation $\rho_{12}^{\text{target}}$ of 0.5, fully dependent and fully independent multiple univariate SGS must be generated first to find coefficients b_{12} and a_{12} in (6.8). Note, however, because the simulation is unconditional coefficient a_{12} should be equal to 0 within acceptable ergodic fluctuation.

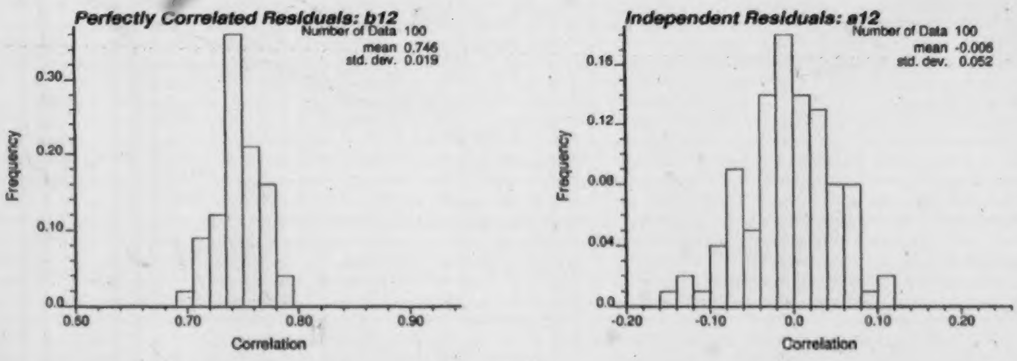
Figure 6.1 shows the histograms of the coefficients b_{12} and a_{12} obtained in 100 fully dependent and fully independent, respectively, multiple univariate SGS realizations. It can be seen from Figure 6.1 that coefficient a_{12} is virtually zero (-0.006) as expected and coefficient $b_{12} = 0.746$. This implies that a correlation coefficient between residuals for generation of Z_1 and Z_2 of

$$\rho_{12} = \frac{\rho_{12}^{\text{target}}}{b_{12}} = \frac{0.5}{0.746} = 0.670 \tag{6.12}$$

need to be applied in the multiple univariate SGS in order to reproduce the target correlation of 0.5. Figure 6.2 shows the distribution of the correlation coefficients between the two random variables under study obtained by multiple univariate Sequential Gaussian Simulation with residual's correlation coefficient equal to 0.670. Figure 6.2 also shows the distribution of the correlation coefficients between realizations for Z_1 and Z_2 obtained by multiple univariate SGS with residual's correlation coefficient equal to 0.5 (target correlation); which is the conventional approach to multiple univariate SGS.

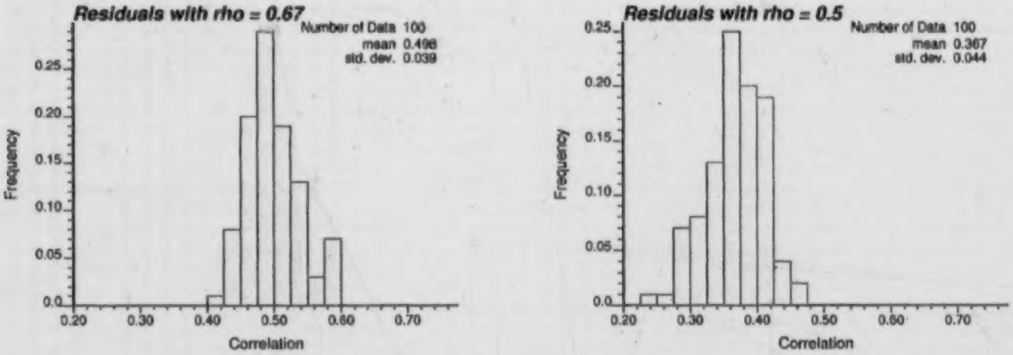
Figure 6.2 shows that the correction approach works perfectly for this example; the target correlation is closely reproduced. The same cannot be said for the conventional approach, where the target correlation is set equal to the target correlation between variables; multiple univariate SGS in this case resulted in a correlation 0.367, which is more than 25% below the target.

Figure 6.3 shows the variogram reproduction for Z_1 and Z_2 obtained in multiple univariate sequential Gaussian simulation with correlation coefficient between residuals fixed at 0.670.



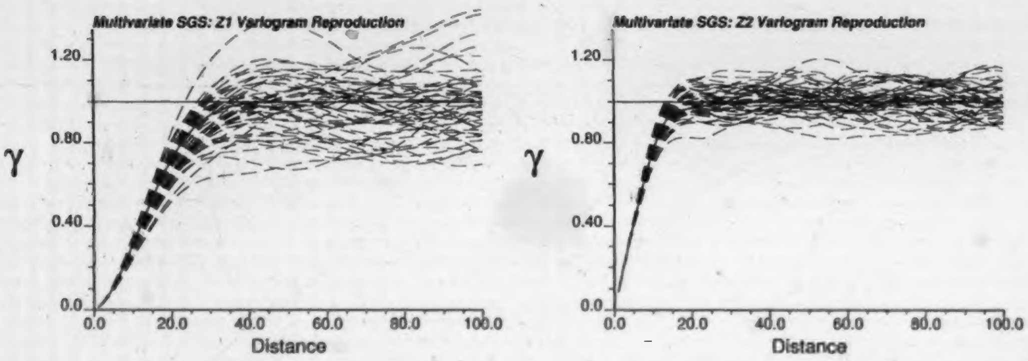
a) b)

Figure 6.1: The histogram of the coefficients b_{12} (a) and a_{12} (b) obtained in 100 fully dependent and fully independent, respectively, multiple univariate SGS.



a) b)

Figure 6.2: Distribution of the correlation coefficients between Z_1 and Z_2 obtained by the corrected multiple univariate SGS (a); and distribution of the correlation coefficients between Z_1 and Z_2 obtained by conventional approach (b).



a) b)

Figure 6.3: The variogram reproduction for Z_1 (a) and Z_2(b) obtained in the corrected multiple univariate SGS.

Note that both variograms are closely reproduced within ergodic fluctuation. This is expected. The correlation between the residuals has no impact on the variogram structure. This is because residuals are independent from location to location. A proof of this is a straightforward extension of the proof for covariance reproduction in the univariate SGS (Deutsch, 2002).

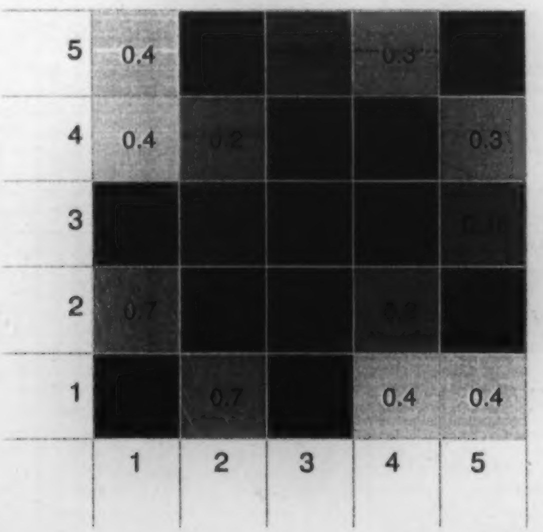
It is also interesting to note that the average largest achievable correlation in multiple univariate SGS between two random variables Z_1 and Z_2 with spatial continuity characterized by $\gamma_{Z_1}(\mathbf{h})$ and $\gamma_{Z_2}(\mathbf{h})$ given by (6.11) is 0.746 (see Figure 6.1). Note that the correlation coefficient of 0.746 does not result in a valid LMC.

6.3.2. Example 2

Let us now consider five standard normal random variables Z_i, $i = 1,\ldots,5$, with the following variograms characterizing their spatial continuity

$$\begin{aligned}
\gamma_{Z_1}(\mathbf{h}) &= 0.1 \cdot Sph_{16}(\mathbf{h}) + 0.9 \cdot Gaus_{32}(\mathbf{h}) \\
\gamma_{Z_2}(\mathbf{h}) &= 0.5 \cdot Exp_{20}(\mathbf{h}) + 0.5 \cdot Sph_{40}(\mathbf{h}) \\
\gamma_{Z_3}(\mathbf{h}) &= 0.3 \cdot Exp_{5}(\mathbf{h}) + 0.7 \cdot Sph_{12}(\mathbf{h}) \quad , \\
\gamma_{Z_4}(\mathbf{h}) &= 0.9 \cdot Sph_{16}(\mathbf{h}) + 0.1 \cdot Gaus_{32}(\mathbf{h}) \\
\gamma_{Z_5}(\mathbf{h}) &= 0.5 \cdot Sph_{16}(\mathbf{h}) + 0.5 \cdot Gaus_{32}(\mathbf{h})
\end{aligned} \tag{6.13}$$

The correlation matrix between these variables is given below:



Now let us consider multiple univariate SGS for simulation of the random variables that reproduce the correlation between variables at lag 0 on the grid of 256 by 256 block of size 1 by 1 unit. The example is divided into 3 parts:

1. multiple univariate SGS for the variables $Z_i, \quad i = 1,2,3;$
2. multiple univariate SGS for the variables $Z_i, \quad i = 1,\ldots,4;$
3. multiple univariate SGS for all five variables, $Z_i, \quad i = 1,\ldots,5.$

The results of the conventional multiple univariate SGS will be compared to the corrected multiple univariate SGS proposed in this thesis.

Note that this 2-D example was chosen small enough to easily analyze the results yet large enough to show realistic variations in the results. In particular, the five variables under study were specifically chosen to exhibit different spatial continuity structures since this is exactly the situation for which proposed correlation correction is developed. The limitation of the proposed correction technique in terms of negative-definiteness of the corrected correction matrix is also illustrated. The conclusions drawn from this example are considered reasonably general.

6.3.2.1. Multiple Univariate SGS for Z_1, Z_2 *and* Z_3

To find the correlation coefficients ρ_{12}, ρ_{13} and ρ_{23} to be used in multiple univariate SGS to reproduce the target correlations of 0.7, -0.2 and -0.5, respectively, fully dependent and fully independent multiple univariate SGS realizations are generated to find the coefficients b_{12}, b_{13}, b_{23}, a_{12}, a_{13} and a_{23} in Equation (6.8). Because the simulation is unconditional, the coefficients $a_{ij}, (i \neq j) = 1,\ldots,3$, should be equal to 0 within acceptable ergodic fluctuation.

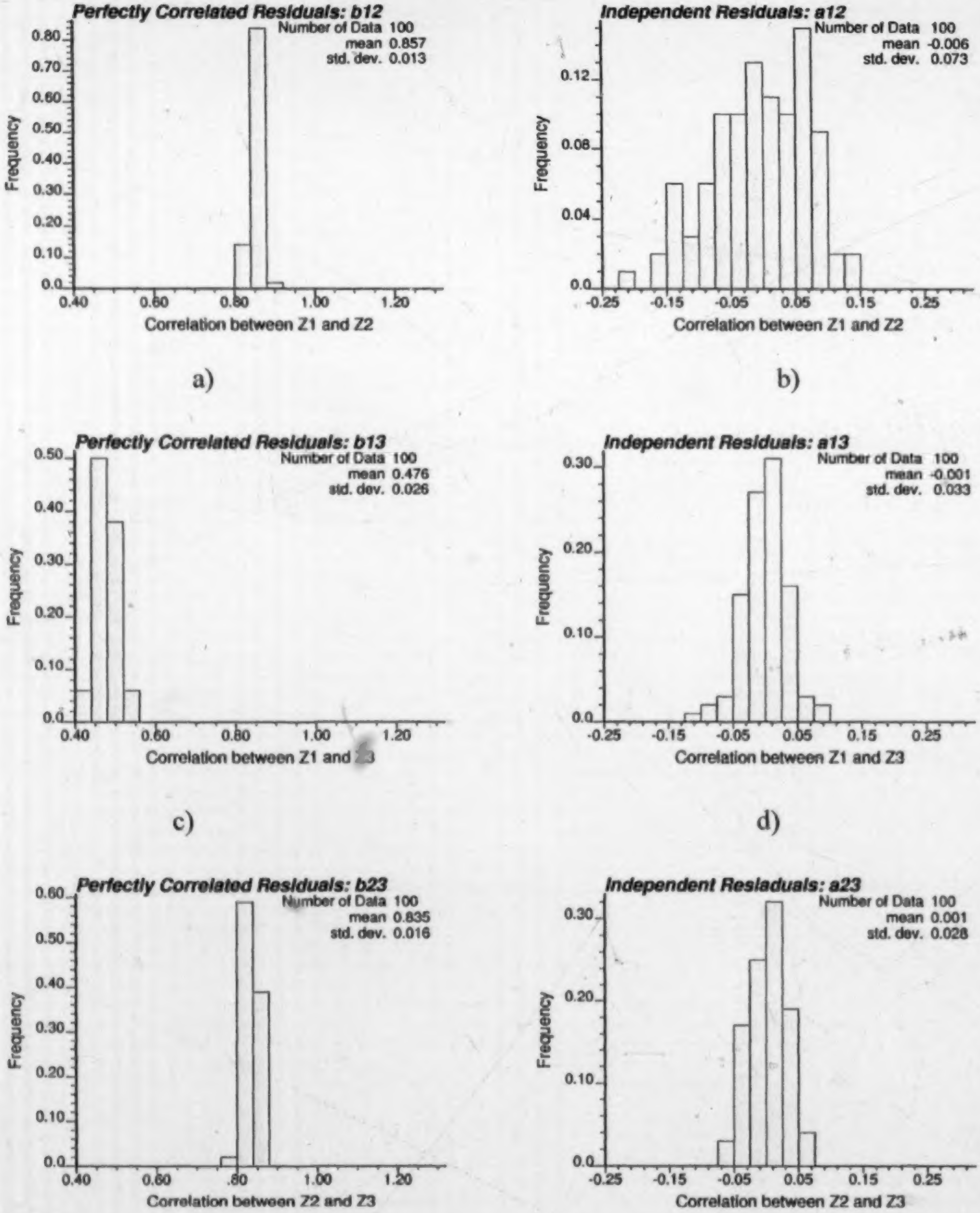
Figure 6.4 shows the distributions of coefficients b_{ij} and $a_{ij}, (i \neq j) = 1,\ldots,3$, obtained by 100 fully dependent and fully independent realizations, respectively. It can be noted from Figure 6.4 that coefficients a_{ij}'s are, as expected, virtually zero; the coefficients b_{ij}'s are given below

$$b_{12} = 0.857; \quad b_{13} = 0.476; \quad b_{23} = 0.835. \tag{6.14}$$

This implies that the following correlation coefficients need to be applied in the multiple univariate SGS

$$\begin{aligned} \rho_{12} &= \frac{0.7}{0.857} = 0.817; \\ \rho_{13} &= \frac{-0.2}{0.476} = -0.420; \\ \rho_{23} &= \frac{-0.5}{0.835} = -0.599. \end{aligned} \tag{6.15}$$

Figure 6.5 shows the distribution of the correlation coefficients ρ_{12}, ρ_{13} and ρ_{23} obtained by the corrected multiple univariate SGS. For comparison, Figure 6.5 also shows the distribution of the correlation coefficients that would be obtained by the conventional approach. The corrected approach results in almost perfect reproduction of the target correlations. The largest absolute mismatch in the correlation coefficients is 0.004. Unfortunately, the same cannot be said about the conventional approach.

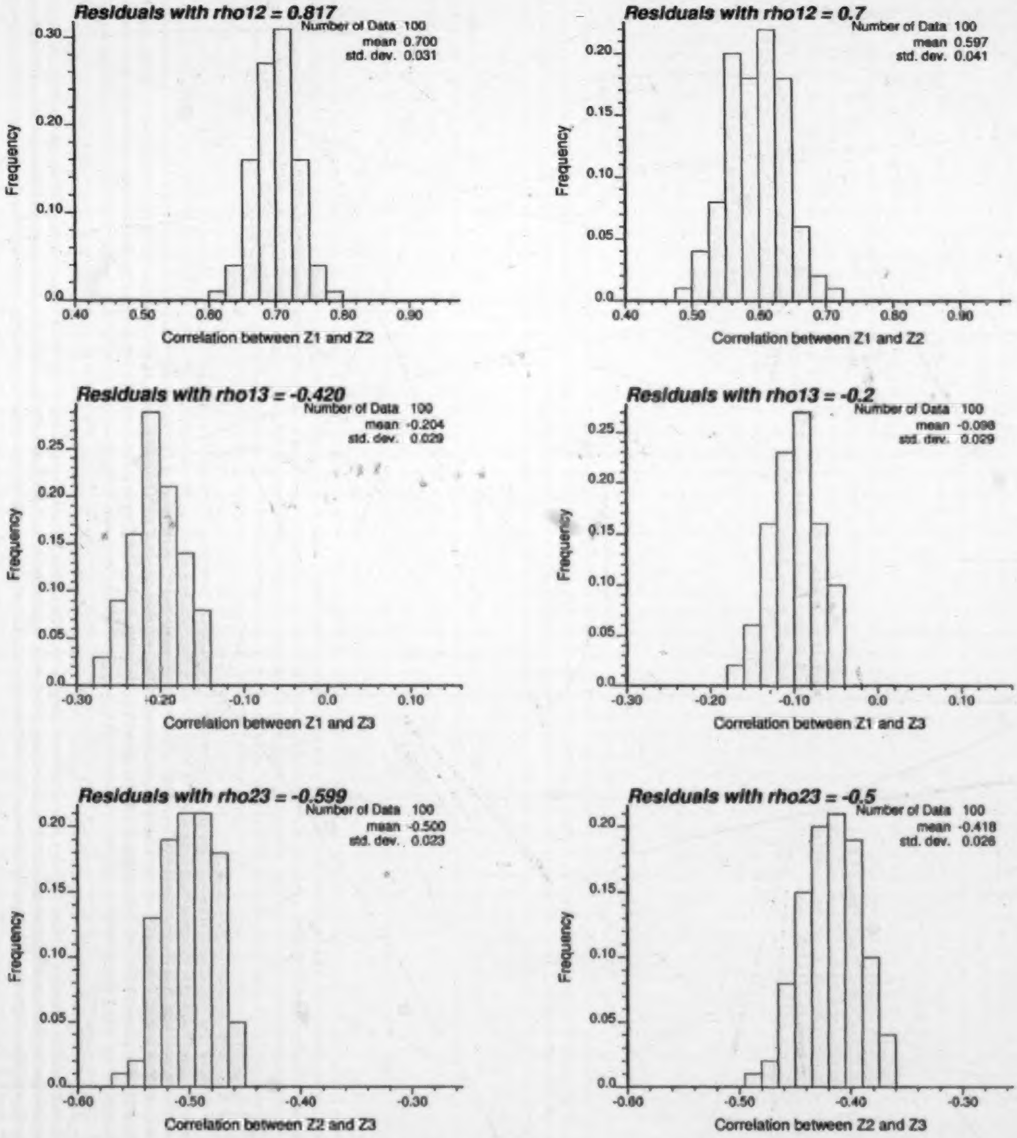


a) b)

c) d)

e) f)

Figure 6.4: Distributions of coefficients b_{12} (a), b_{13} (c), b_{23} (e), a_{12} (b), a_{13} (d), and a_{23} (f) obtained by 100 fully dependent and fully independent, respectively, multiple univariate SGS.



a) b)

Figure 6.5: Distribution of the correlation coefficients ρ_{12}, ρ_{13} and ρ_{23} obtained by the corrected multiple univariate SGS (a) and by conventional approach (b).

The largest absolute mismatch in the correlation coefficients in the conventional approach is more than 0.1, which is quite significant.

Figure 6.6 shows the variogram reproduction for all three variables obtained in the multiple univariate SGS with corrected correlation matrix in Equation (6.15). All variograms are nicely reproduced within ergodic fluctuation.

6.3.2.2. Multiple Univariate SGS for Z_1, Z_2, Z_3*, and* Z_4

To find the correlation coefficients ρ_{12}, ρ_{13}, ρ_{14}, ρ_{23}, ρ_{24} and ρ_{34} to be used in corrected multiple univariate SGS, fully dependent and fully independent multiple univariate SGS are generated first to find the coefficients b_{ij}, a_{ij}, $(i \neq j) = 1,\ldots,4$, in Equation (6.8). Once again, because the simulation is unconditional, the coefficients a_{ij}, $(i \neq j) = 1,\ldots,4$, should be equal to 0 within acceptable ergodic fluctuation.

The following is a summary of the results for coefficients b_{ij}, a_{ij}, $(i \neq j) = 1,\ldots,4$,

$$\begin{aligned} &a_{12} = -0.005;\ a_{13} = -0.002;\ a_{14} = 0.009;\ a_{23} = 0.005;\ a_{24} = 0.003;\ a_{34} = -0.004; \\ &b_{12} = 0.858;\ b_{13} = 0.478;\ b_{14} = 0.744;\ b_{23} = 0.834;\ b_{24} = 0.949;\ b_{34} = 0.895; \end{aligned} \qquad (6.16)$$

This implies that the following correlation coefficients need to be used in multiple univariate SGS:

$$\begin{aligned} &\rho_{12} = \frac{0.7}{0.858} = 0.816;\quad \rho_{13} = \frac{-0.2}{0.476} = -0.420;\quad \rho_{14} = \frac{0.4}{0.744} = 0.538; \\ &\rho_{23} = \frac{-0.5}{0.834} = -0.600;\quad \rho_{24} = \frac{0.2}{0.949} = 0.211;\quad \rho_{34} = \frac{-0.01}{0.895} = 0.011. \end{aligned} \qquad (6.17)$$

Note that the corrected correlation coefficients corresponding to the correlations between the first three random variables are exactly the same. This observation confirms the linear relationship in the correlation and independence of the solutions for the correlation coefficients.

Figure 6.7 shows the correlation matrix between variables reproduced by newly proposed multiple univariate SGS. For comparison, Figure 6.7 also shows the correlation matrix reproduced by the conventional approach. The mismatch in the results for the correlation matrix obtained by the two approaches to multiple univariate SGS are shown in Figure 6.8. Note that the maximum absolute mismatch in the correlation obtained by the corrected approach is 0.005, while in the conventional approach it is 0.105.

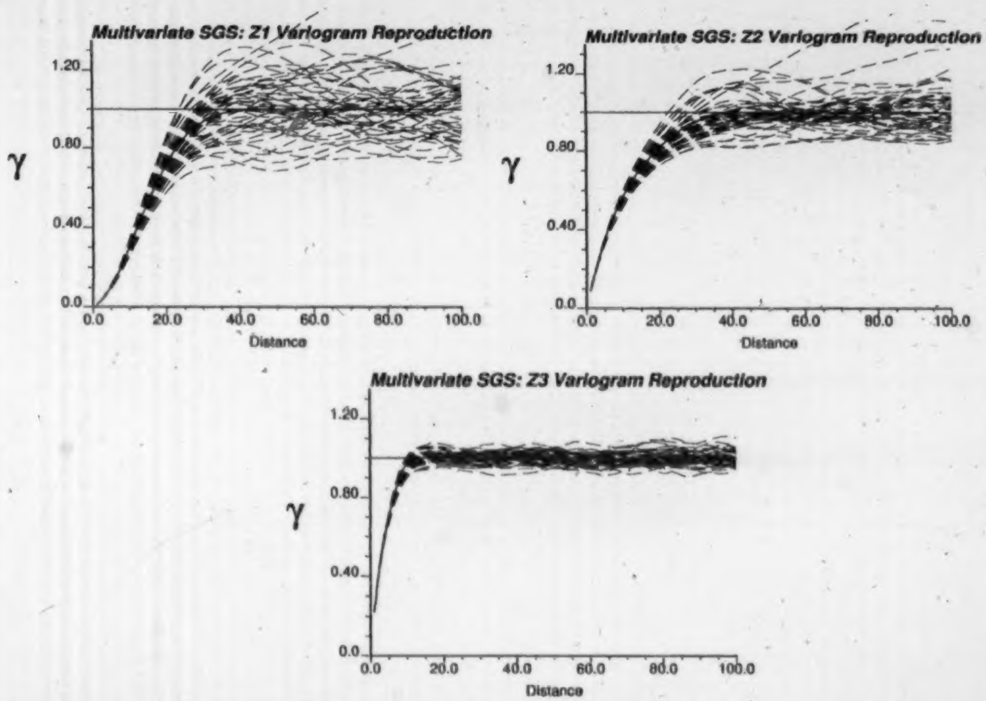
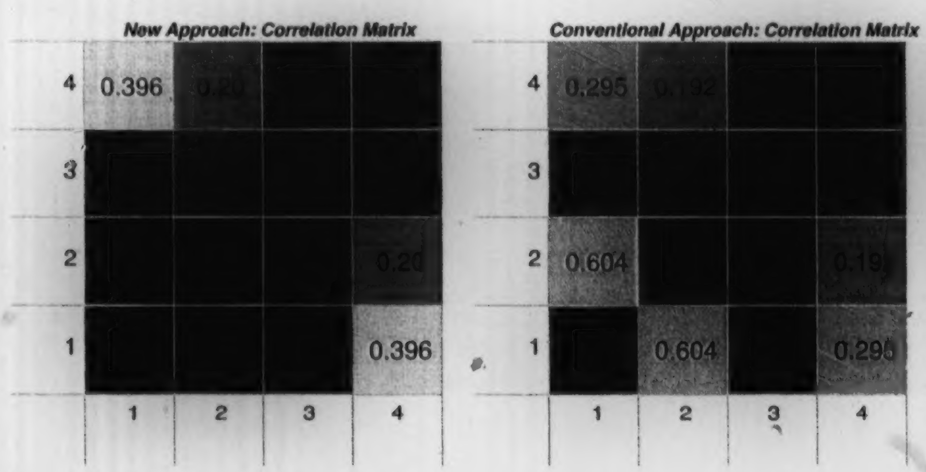


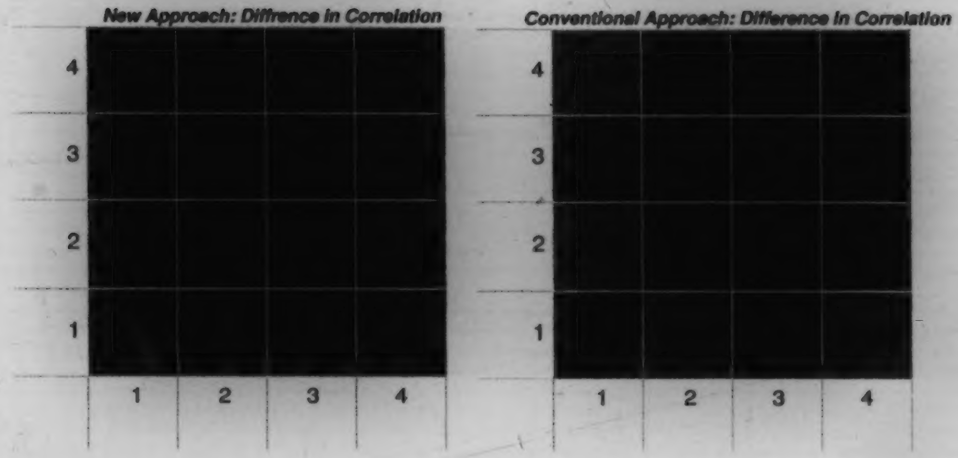
Figure 6.6: Variogram reproduction for Z_1, Z_2, and Z_3 obtained in the corrected multiple univariate SGS.



a)

b)

Figure 6.7: Correlation matrix between Z_1, Z_2, Z_3, and Z_4 reproduced by the corrected multiple univariate SGS (a) and by conventional approach (b).



New Approach: Diffrence in Correlation

Conventional Approach: Difference in Correlation

a) b)

Figure 6.8: The mismatch in the reproduced correlation matrix between Z_1, Z_2, Z_3, and Z_4 obtained by the corrected multiple univariate SGS (a) and by conventional approach (b).

Figure 6.9 shows the variogram reproduction for all four variables obtained in the multiple univariate SGS with the corrected correlation matrix. All variograms, as expected, are nicely reproduced within ergodic fluctuation.

6.3.2.3. Multiple Univariate SGS for All Five Variables

The same procedure is applied to find the coefficients a_{ij}, b_{ij}, $(i \neq j) = 1, \ldots, 5$, in Equation (6.8). The following is a summary of the results for the b_{ij}'s coefficients:

$$\begin{aligned} &b_{12} = 0.857; \quad b_{13} = 0.476; \quad b_{14} = 0.742; \quad b_{15} = 0.873; \quad b_{23} = 0.834; \\ &b_{24} = 0.949; \quad b_{25} = 0.999; \quad b_{34} = 0.895; \quad b_{35} = 0.818; \quad b_{45} = 0.943; \end{aligned} \tag{6.18}$$

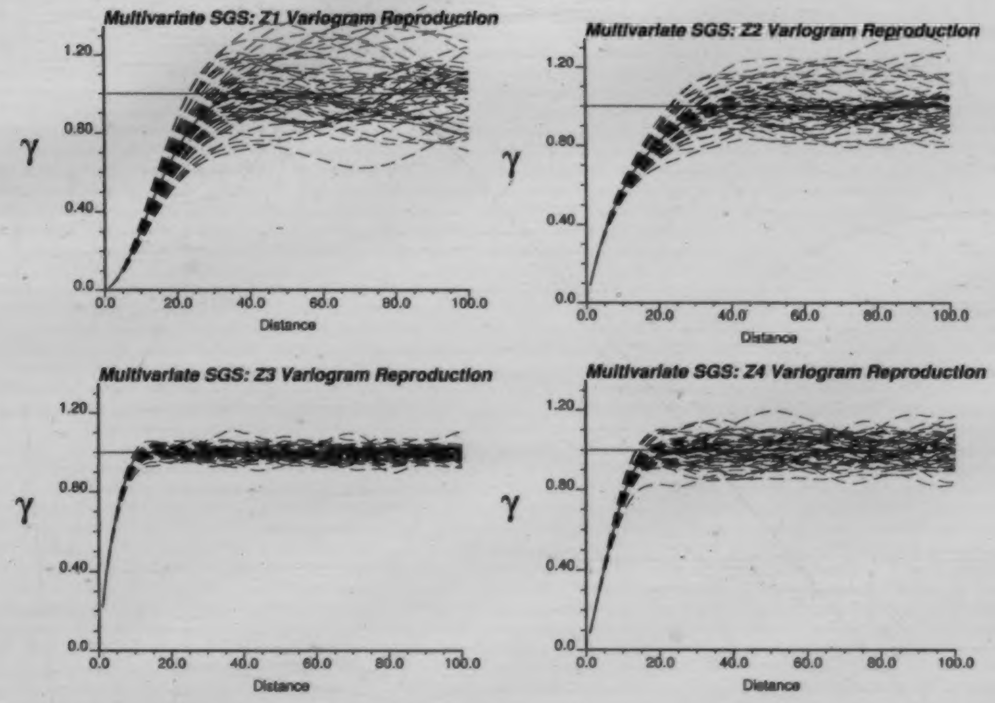


Figure 6.9: Variogram reproduction for Z_1, Z_2, Z_3 and Z_4 obtained in the corrected multiple univariate SGS.

This implies that the following correlation coefficients:

$$\begin{gathered}
\rho_{12}=\frac{0.7}{0.857}=0.817;\quad \rho_{13}=\frac{-0.2}{0.476}=-0.420;\quad \rho_{14}=\frac{0.4}{0.742}=0.539;\\
\rho_{15}=\frac{0.4}{0.873}=0.458;\quad \rho_{23}=\frac{-0.5}{0.834}=-0.600;\quad \rho_{24}=\frac{0.2}{0.949}=0.211;\\
\rho_{25}=\frac{-0.2}{0.999}=-0.2;\quad \rho_{34}=\frac{-0.01}{0.895}=0.011;\quad \rho_{35}=\frac{0.15}{0.818}=0.183;\\
\rho_{45}=\frac{0.3}{0.943}=0.318.
\end{gathered} \tag{6.19}$$

As before, the corrected correlation coefficients corresponding to correlation between first four random variables remain unchanged.

Note, however, that the combined matrix of correlation coefficients is not positive definite. The smallest eigenvalue of the corrected correlation matrix is -0.0434. Thus, a correction must be applied. Specifically, to make the correlation matrix positive-semi-definite, the off-diagonal elements in this matrix should be standardized by (1 – smallest eigenvalue + ε), where ε is a very small value, e.g., 0.00001. In our case the off-diagonal elements are standardized by 1.0434. As a result, the positive-definiteness correction is minimal.

The following matrix shows the input correlation matrix to the multiple univariate SGS (the correlation matrix was made positive definite)

$$\boldsymbol{\rho} = \begin{bmatrix} 1.0000 & 0.7829 & -0.4028 & 0.5167 & 0.4391 \\ 0.7829 & 1.0000 & -0.5748 & 0.2020 & -0.1919 \\ -0.4028 & -0.5748 & 1.0000 & -0.0107 & 0.1757 \\ 0.5167 & 0.2020 & -0.0107 & 1.0000 & 0.3048 \\ 0.4391 & -0.1919 & 0.1757 & 0.3048 & 1.0000 \end{bmatrix} \quad (6.20)$$

Figure 6.10 shows the correlation matrix between variables reproduced by corrected multiple univariate SGS. For comparison, Figure 6.10 also shows the correlation matrix from the conventional approach. The mismatch in the results for correlation obtained by the two approaches to multiple univariate SGS are shown in Figure 6.11. Note that the maximum absolute mismatch in the correlations obtained using the corrected approach is 0.028, while in the conventional approach gives a 0.107 difference. The mismatch in the reproduction of the target correlation coefficients obtained via corrected approach is slightly higher than before. This is connected to the correction of the input correlation matrix to make it positive definite.

6.4. Conditional Multiple Univariate SGS: Example

Let us consider the same data as in Section 5.5.2. Figure 6.12 shows locations of 20 primary data in the study domain of size 100 by 100 blocks of size 1 by 1 unit; primary data distribution, the crossplot between primary data and collocated secondary data and the distribution of the secondary data collocated to primary. All data are in Gaussian units.

The following linear model of coregionalization describes the joint continuity of the primary and secondary data:

$$\begin{aligned}\gamma_{YY}(\mathbf{h}) &= 0.3\cdot Exp_{\substack{a_1=10\\a_2=20}}(\mathbf{h}) + 0.7\cdot Sph_{\substack{a_1=20\\a_2=40}}(\mathbf{h})\\ \gamma_{YZ}(\mathbf{h}) &= 0.45\cdot Exp_{\substack{a_1=10\\a_2=20}}(\mathbf{h}) + 0.35\cdot Sph_{\substack{a_1=20\\a_2=40}}(\mathbf{h}),\\ \gamma_{ZZ}(\mathbf{h}) &= 0.8\cdot Exp_{\substack{a_1=10\\a_2=20}}(\mathbf{h}) + 0.2\cdot Sph_{\substack{a_1=20\\a_2=40}}(\mathbf{h})\end{aligned} \tag{6.21}$$

Now, let us consider simulating the primary and secondary random variables using corrected multiple univariate SGS. Figure 6.13 shows the distributions of coefficients a_{12} and b_{12} obtained using the approach outlined in Section 6.2,

$$a_{12} = 0.087 \quad \text{and} \quad b_{12} = 0.842.$$

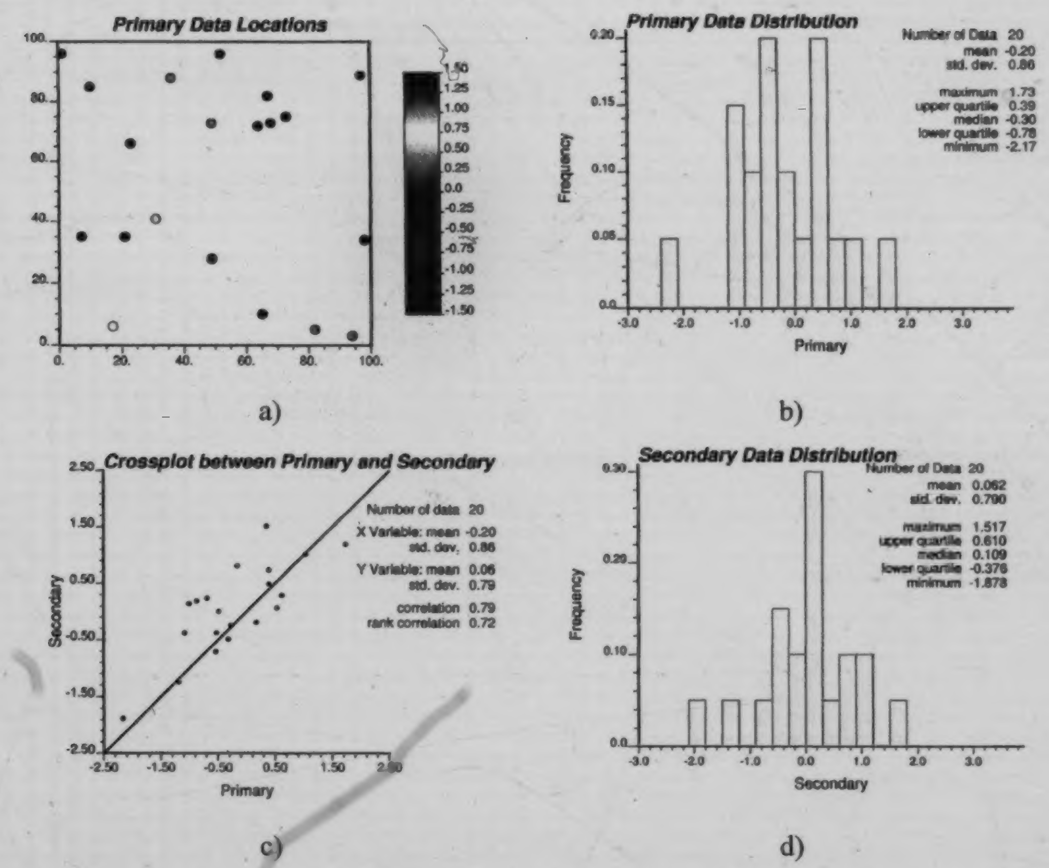
Thus a correlation coefficient between residuals of

$$\rho_{12} = \frac{\rho_{12}^{\text{target}} - a_{12}}{b_{12}} = \frac{0.8-0.087}{0.842} = 0.847 \tag{6.22}$$

needs to be applied in the multiple univariate SGS in order to reproduce the target correlation of 0.8.

Figure 6.14 shows the distribution of the correlation coefficients between the primary and secondary random variables obtained by multiple univariate sequential Gaussian simulation with residual's correlation coefficient given in (6.22). Note from Figure 6.14 that the target correlation is nicely reproduced.

Figure 6.15 shows the variogram reproduction for the primary and secondary random variables obtained in multiple univariate sequential Gaussian simulation with correlation coefficient between residuals fixed at 0.847.



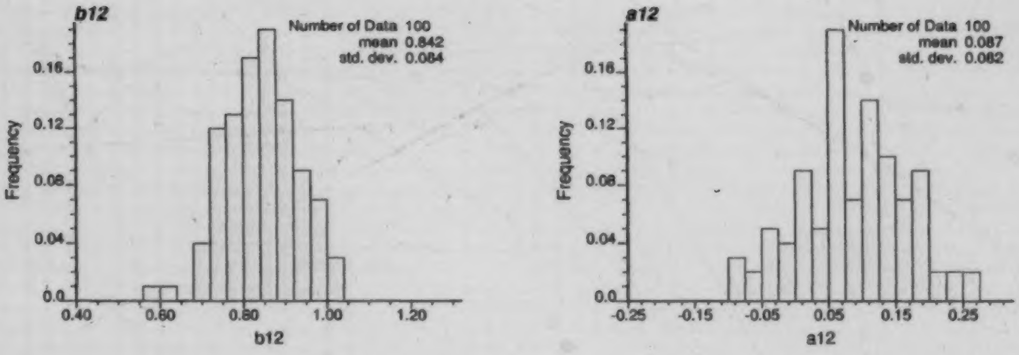
a)

b)

c)

d)

Figure 6.12: Locations of the 20 primary data (a) and their distribution (b); the crossplot between primary data and collocated secondary data (c) and the distribution of the secondary data (d). The data are in Gaussian units.



a) b)

Figure 6.13: Distributions of coefficients b_{12} (a) and a_{12} (b) obtained in conditional multiple univariate SGS example.

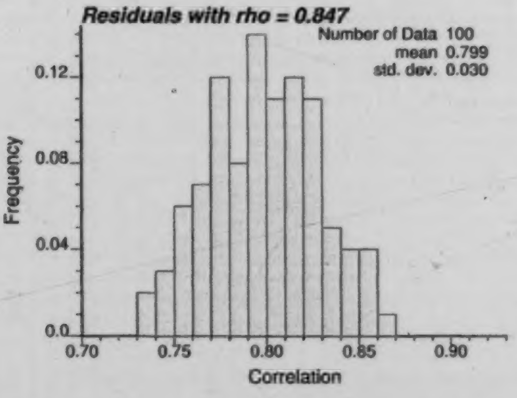
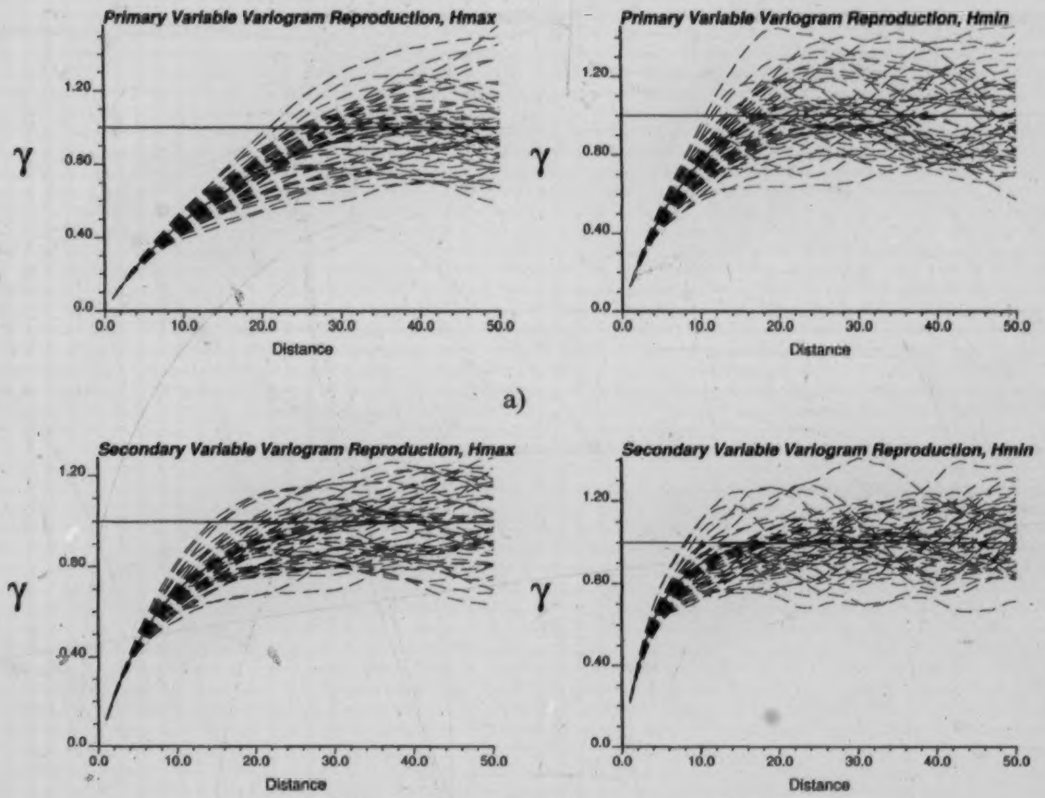


Figure 6.14: Distribution of the correlation coefficients between primary and secondary random variables obtained by the corrected multiple univariate SGS.



a)

b)

Figure 6.15: Variogram reproduction in the direction of major and minor continuity for primary (a) and secondary (b) random variables obtained in the corrected multiple univariate SGS.

Note that variogram reproduction is acceptable (that is, average variogram reproduced by multiple univariate SGS is virtually the same as the target) within ergodic fluctuation. Also, it is interesting to note that the cross variogram between variables given in Equation (6.21) is well reproduced, see Figure 6.16.

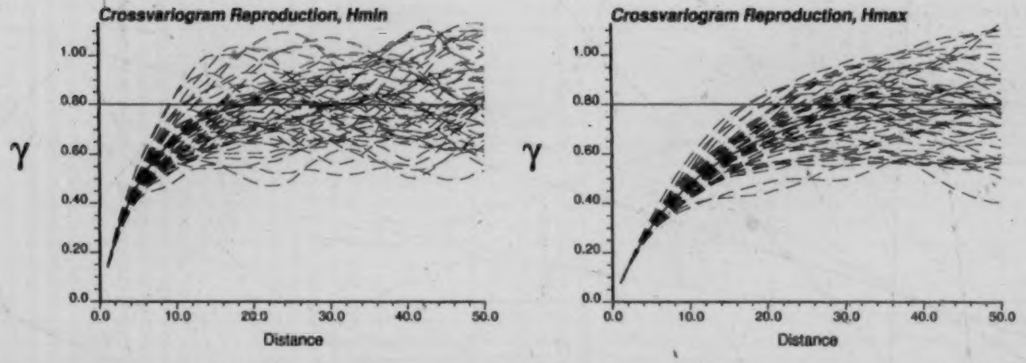


Figure 6.16: Cross variogram reproduction in the direction of major and minor continuity obtained in the corrected multiple univariate SGS.

6.5. Local Correlation

With a reasonable practical effort, the correction proposed for multiple univariate SGS can be localized. That is, multiple univariate unconditional SGS with locally varying correlated residuals $\mathbf{R}(\mathbf{u})^T = [R_1(\mathbf{u}),\ldots,R_N(\mathbf{u})]^T$ can be developed. The procedure for finding the prescribed locally varying correlation matrix $\boldsymbol{\rho}(\mathbf{u})$ of the residuals to reproduce the target correlation-covariance matrix for the random variables to be simulated, $\boldsymbol{\rho}^{\text{target}}$, is as follows.

The residuals $\mathbf{R}(\mathbf{u})^T = [R_1(\mathbf{u}),\ldots,R_N(\mathbf{u})]^T$ are assumed to be independent from location to location, that is,

$$\rho(R_i(\mathbf{u}_k), R_j(\mathbf{u}_l)) = Cov(R_i(\mathbf{u}_k), R_j(\mathbf{u}_l)) = 0, \quad \forall i, j = 1,\ldots,N, \quad \forall k \neq l; \qquad (6.23)$$

but correlated at the same location with locally varying correlation-covariance matrix $\boldsymbol{\rho}(\mathbf{u})$, that is,

$$\rho(R_i(\mathbf{u}_k), R_j(\mathbf{u}_k)) = Cov(R_i(\mathbf{u}_k), R_j(\mathbf{u}_k)) = \rho_{ij}(\mathbf{u}), \quad \forall i, j = 1,\ldots,N, \quad \forall k. \qquad (6.24)$$

Let us now calculate the correlation-covariance between two simulated values for variables Z_i and Z_j, respectively, at arbitrary location $\mathbf{u}$ in the study domain,

$i, j \in \{1, \ldots, N\}$. These simulated values $Z_i(\mathbf{u})$ and $Z_j(\mathbf{u})$ in sequential Gaussian simulation are given by

$$\begin{aligned} Z_i(\mathbf{u}) &= m_{SK,i}(\mathbf{u}) + \sigma_{SK,i}(\mathbf{u}) R_i(\mathbf{u}), \\ Z_j(\mathbf{u}) &= m_{SK,j}(\mathbf{u}) + \sigma_{SK,j}(\mathbf{u}) R_j(\mathbf{u}), \end{aligned} \tag{6.25}$$

then due to independence between the vector of residuals $[R_i(\mathbf{u})\ R_j(\mathbf{u})]^T$ and the vector of Simple Kriging means $[m_{SK,i}(\mathbf{u})\ m_{SK,j}(\mathbf{u})]^T$,

$$\begin{aligned} \rho(Z_i(\mathbf{u}), Z_j(\mathbf{u})) &= Cov(Z_i(\mathbf{u}), Z_j(\mathbf{u})) \\ &= Cov(m_{SK,i}(\mathbf{u}), m_{SK,j}(\mathbf{u})) + \sigma_{SK,i}(\mathbf{u})\sigma_{SK,j}(\mathbf{u}) Cov(R_i(\mathbf{u}), R_j(\mathbf{u})) \\ &= Cov(m_{SK,i}(\mathbf{u}), m_{SK,j}(\mathbf{u})) + \sigma_{SK,i}(\mathbf{u})\sigma_{SK,j}(\mathbf{u}) \rho_{ij}(\mathbf{u}). \end{aligned} \tag{6.26}$$

Note that

$$\begin{aligned} m_{SK,i}(\mathbf{u}) &= \sum_{k=1}^{n_i(\mathbf{u})} \lambda^i_{SK,k} Z_i(\mathbf{u}_k), \\ m_{SK,j}(\mathbf{u}) &= \sum_{l=1}^{n_j(\mathbf{u})} \lambda^j_{SK,k} Z_j(\mathbf{u}_l), \end{aligned} \tag{6.27}$$

where $Z_i(\mathbf{u}_k)$, $k = 1, \ldots, n_i(\mathbf{u})$, and $Z_j(\mathbf{u}_k)$, $l = 1, \ldots, n_j(\mathbf{u})$ denote the $n_i(\mathbf{u})$ and $n_j(\mathbf{u})$ closest simulated nodes for variables Z_i and Z_j, respectively; $\lambda^i_{SK,k}$, $k = 1, \ldots, n_i(\mathbf{u})$, and $\lambda^i_{SK,l}$, $l = 1, \ldots, n_j(\mathbf{u})$, denote the Simple Kriging weights obtained for location $\mathbf{u}$ when estimating variables Z_i and Z_j, respectively.

Because the simulation is unconditional the Simple Kriging means given in (6.27) can be rewritten as

$$\begin{aligned} m_{SK,i}(\mathbf{u}) &= \sum_{k=1}^{N_i(\mathbf{u})} \mu^i_k R_i(\mathbf{u}_k), \\ m_{SK,j}(\mathbf{u}) &= \sum_{l=1}^{N_j(\mathbf{u})} \mu^j_k R_j(\mathbf{u}_l), \end{aligned} \tag{6.28}$$

where $R_i(\mathbf{u}_k)$, $k = 1, \ldots, N_i(\mathbf{u})$, and $R_j(\mathbf{u}_k)$, $l = 1, \ldots, N_j(\mathbf{u})$ denote the Gaussian residuals generated for calculation of the $n_i(\mathbf{u})$ and $n_j(\mathbf{u})$ closest simulated nodes to estimation location $\mathbf{u}$ for variables Z_i and Z_j, respectively; μ^i_k, $k = 1, \ldots, N_i(\mathbf{u})$, and μ^i_l, $l = 1, \ldots, N_j(\mathbf{u})$, denote the weights given to these residuals.

Then,

$$\begin{aligned} Cov(m_{SK,i}(\mathbf{u}), m_{SK,j}(\mathbf{u})) &= Cov\left(\sum_{k=1}^{N_i(\mathbf{u})} \mu_k^i R_i(\mathbf{u}_k), \sum_{l=1}^{N_j(\mathbf{u})} \mu_k^j R_j(\mathbf{u}_l) \right) \\ &= \sum_{k=1}^{N_i(\mathbf{u})} \sum_{l=1}^{N_j(\mathbf{u})} \mu_k^i \mu_k^j Cov(R_i(\mathbf{u}_k), R_j(\mathbf{u}_l)) = \sum_{s=1}^{N(\mathbf{u})} \widetilde{\mu}_s^j \widetilde{\mu}_s^i \rho_{ij}(\mathbf{u}_s), \end{aligned} \tag{6.29}$$

where $N(\mathbf{u})$ denotes the number of location with residuals common to both random variables; $\widetilde{\mu}_s^i, \widetilde{\mu}_s^i$ denote the residual weights assigned to location with residuals common to both random variables; $\rho_{ij}(\mathbf{u}_s)$ denotes the correlation between residuals at location with residuals common to both random variables $s = 1, \ldots, N(\mathbf{u})$.

Moreover, because we aim at

$$\rho(Z_i(\mathbf{u}), Z_j(\mathbf{u})) = \rho_{ij}^{\text{target}}, \tag{6.30}$$

the following equality must hold

$$\rho_{ij}^{\text{target}} = \sum_{s=1}^{N(\mathbf{u})} \widetilde{\mu}_s^j \widetilde{\mu}_s^i \rho_{ij}(\mathbf{u}_s) + \sigma_{SK,i}(\mathbf{u}) \sigma_{SK,j}(\mathbf{u}) \rho_{ij}(\mathbf{u}). \tag{6.31}$$

Thus, in order for the multiple univariate unconditional Sequential Gaussian Simulation to honor the locally varying correlation matrix between random variables at lag 0, $\boldsymbol{\rho}^{\text{target}}(\mathbf{u})$, the residuals with the following correlation structure need to be generated locally for each simulation location $\mathbf{u}$

$$\rho_{ij}(\mathbf{u}) = \begin{cases} \dfrac{\rho_{ij}^{\text{target}} - \sum_{s=1}^{N(\mathbf{u})} \widetilde{\mu}_s^j \widetilde{\mu}_s^i \rho_{ij}(\mathbf{u}_s)}{\sigma_{SK,i}(\mathbf{u}) \sigma_{SK,j}(\mathbf{u})}, & if \left| \dfrac{\rho_{ij}^{\text{target}} - \sum_{s=1}^{N(\mathbf{u})} \widetilde{\mu}_s^j \widetilde{\mu}_s^i \rho_{ij}(\mathbf{u}_s)}{\sigma_{SK,i}(\mathbf{u}) \sigma_{SK,j}(\mathbf{u})} \right| \le 1, \\ 1 \quad \text{or} \quad -1, & \text{otherwise}; \end{cases} \tag{6.32}$$

$$\rho_{ii}(\mathbf{u}) = 1, \quad \forall i = 1, \ldots, N.$$

for any $(i \neq j) = 1, \ldots, N$.

An unfortunate feature of the correction (6.32) (as is of global correction (6.10)) is that combined matrix of the locally varying correlations $\boldsymbol{\rho}(\mathbf{u})$ to be used in the multiple univariate SGS to reproduce the locally varying target correlation $\boldsymbol{\rho}^{\text{target}}(\mathbf{u})$ is not necessarily positive definite. Therefore if matrix (6.32) is not positive definite, a positive definiteness correction to this matrix must be applied first at the location of non-positive-definiteness, then it can be used in multiple univariate SGS. The only situation where matrix (6.32) is known to be positive definite at any estimation location **u** is in the case of multiple univariate unconditional Gaussian simulation with only two random variables.

6.6. Discussion

Multiple univariate sequential Gaussian simulation reproducing correlation between variables at lag 0 represents a neat alternative to sequential Gaussian simulation with intrinsic collocated cokriging and sequential Gaussian simulation with simple cokriging in the case when the secondary data is not exhaustively sampled or simulation is aimed at generating multiple primary variables.

The approach with correlated residuals is simple; it ensures reproduction of the mean variance and target direct variograms. Moreover, sometimes even the cross variograms between variables are reproduced quite well. The newly proposed approach supplements other well founded techniques for reproducing lag 0 correlation matrix such as kriging or cokriging of principal components or minmax autocorrelated factors. The advantage of the multiple univariate sequential Gaussian simulation reproducing correlation between variables at lag 0 over these other techniques is that it ensures reproduction of the variogram models and covariances.

A drawback of the proposed approach is possible non-positive definiteness of the correlation matrices for the residuals to be used in multiple univariate Sequential Gaussian Simulation to honor correlation between variables at lag 0.

The global correction for the correlation matrix of residuals is "global" over the entire domain under consideration. The correction ought to be different depending on the location of nearby conditioning data and the size of the study domain. That is, the a's and b's are calculated in expected value over all locations. This is a practical approach with limited consequences.

CHAPTER 7

Discussion and Conclusions

Despite the widespread used of geostatistical techniques, there are unresolved problems with commonly applied geostatistical estimation and simulation methods. These problems include variance inflation, biased reproduction of multivariate statistics, biased estimation or difficult implementation. This thesis develops theoretically sound and/or practical methodologies for improved geostatistical estimation and simulation. Improved methodologies have been proposed for the following five important longstanding problems in geostatistics:

1) String effect of kriging;
2) Variance inflation of collocated simple cokriging;
3) Accounting for multiple secondary attributes in collocated cokriging;
4) Obtaining a combined ('best') measure of local uncertainty; and
5) Correcting multiple univariate sequential Gaussian simulation with correlated residuals.

The main results and conclusions that can be extracted from this research related to these problems are summarized below.

7.1. Summary

7.1.1. Correcting the String Effect

Kriging assigns unreasonably large weights to the end samples in strings of data. Strings of data are often observed in mining and petroleum applications where the data are collected along drillholes or wells. The weights are theoretically valid and optimal in a stationary multivariate Gaussian setting; however, the large weights applied to the boundary samples ('string effect') can cause problems in practical estimation and simulation especially when the domain under study is non-stationary, that is, when the data at the ends of the finite string the data values are higher or lower than in the middle of the string.

A number of ad-hoc solutions have been developed for the string effect of kriging. The most accepted are to limit the number of data from a drillhole, wrap the string, or to extend the string and discard weights at the end samples. These ad-hoc approaches aim at fixing the string effect either by changing the data configuration or the covariance function. None of them provides an authomatic constrained solution with a well defined measure of optimality.

Two new approaches for correcting thestring effect are proposed. These methods are distance constrained kriging and finite domain kriging. Distance constrained kriging method corrects the string effect by constraining the kriging weights to have a certain distance influence structure. That is, the weights are ordered with respect to distance from the unsampled location: the closest data in the string is constrained to receive the largest weight; the second closest data is constrained to receive the second largest weight and so on. The data in the string located furthest from the estimation location is assigned the smallest weight. The optimal distance constrained kriging estimator is obtained by minimizing the estimation variance subject to these distance constraints.

The distance constrained kriging method for estimation in a finite domain possesses several important properties. Distance constrained kriging estimator is linear unbiased and exact. Similar to inverse distance interpolation, the distance constrained kriging orders weights according to the distance. However, as in both ordinary and simple kriging, the magnitude of the weights accounts for the spatial continuity of the variable under study through the variogram model. Like traditional kriging, distance constrained kriging takes into account the redundancy of data (defined by data-to-data covariance) in the string and closeness of the data in the string to the estimation location (defined by data-to-estimation point covariance).

A second method for correcting the string effect of kriging, that is, finite domain kriging, is based on kriging with a successively larger number of data. The total number of relevant neighbor data (*n*) are established first, then *n* kriging matrices are found and *n* kriging systems are solved subsequently to arrive at kriging weights that allow all data in the string to be considered end samples. The first kriging matrix has only the closest single data value (*n*-1 weights for the rest of the data are set to zero), the second has the two closest data (*n*-2 weights are set to zero), and so on. The final matrix is computed based on all *n* data. Finite domain kriging estimator is obtained as an average (or expected value) of the optimal kriging estimators for different search neighborhoods.

The finite domain kriging estimator is also unbiased and exact. It has similar characteristics to both inverse distance and kriging. However, on the contrary to distance constrained kriging, finite domain kriging not only removes large weights assigned for end samples (string effect), but also ensures only minimal correction of the kriging weights. No constraints on the weights that may lead to suboptimal estimation are introduced. Furthermore, the finite domain kriging is convergent. That is, estimation of a particular location of interest based on a very long ('infinite') string of data is virtually equivalent to estimation of this particular location based on a long portion of this string.

Both newly proposed methods for correcting the string effect are applicable when the data at the end of strings of data are somehow anomalous, for example, with thin deposits with vertical trends. A small case study in Chapter 3 shows that finite domain kriging and distance constrained kriging provide improved estimates. In this particular case study, finite domain kriging was shown to result in stronger correction; however, distance constrained kriging performed better in jackknife validation. The situation could be different in different areas or with different data sets. Note also that when applying "string effect" corrected kriging approaches a special caution must be taken. This is because when the sudy domain is stationary and the estimated covariance of the data closely approaches the true covariance model, traditional kriging approaches will provide by construction the best estimates. In this case correcting the string problem to account for non-stationarity has the consequence of providing sub-optimal weights and estimates.

7.1.2. Uncertainty as the Overlap of Distributions

An important task in modern geostatistics is the assessment and quantification of resource and reserve uncertainty. This uncertainty is required for risk assessment and decision-making. There are many different methods/interpolation techniques to build models of resource uncertainty including kriging, cokriging and inverse distance, etc. Each method leads to different results and is good in different senses, e.g., simplicity, robustness, reliability, flexibility, geological realism or statistical accuracy. Because there is no unique technique that can be regarded as the best one for any area or spatial and temporal scale in terms of its accuracy and profitability, researchers need to combine spatial predictions and associated models of uncertainty obtained by different models to get a best result given available information. There are many different ways to combine local estimators of uncertainty. These include combing local uncertainties by a linear weighted average or, if there is belief that estimators are in some sort of disagreement, then multi-modal local uncertainties calculated as the maximum of the local uncertainties obtained by different estimation techniques can be used.

In Chapter 4 another interesting approach for combining alternate local conditional distributions is proposed. This approach, referred to as overlap uncertainty, forces a compromise between different estimators by calculating the local uncertainty as the minimum of the local uncertainties obtained by these estimation techniques (scaled to 1). The overlap uncertainty approach is illustrated using overlap between simple kriging and inverse distance methods. Both simulated and real case studies are considered. It is shown that overlap uncertainty estimator can result in significantly narrower intervals for the local uncertainty. Different local conditional distributions could result in wider, not narrower, intervals of uncertainty. This would represent the fact that estimators are in significant disagreement at the unsampled location.

Moreover, it is noted that the results for the overlap uncertainty estimator are sensitive to the estimators used for overlap of uncertainty. Before applying the proposed overlap technique, one should always check whether the local distributions are accurate.

Also it must be noted that the overlap uncertainty approach is only aimed at combing alternate local conditional distributions of continuous variables with an infinite support. In should not be applied to discrete variables or continuous variables with a finite support.

7.1.3. Intrinsic Collocated Cokriging

Simulation is a powerful tool for modeling variables that cannot be described deterministically due to their inherent complexity. A great number of methods have been developed for joint simulation of dependent random variables. One of the most popular and simplest methods for modeling primary variable based on extensively sampled secondary information is sequential simulation with collocated simple cokriging. The method is widely used because of its simplicity; the correlation coefficient between the primary variable being modeled and the secondary data is the only statistic required to integrate the secondary data in estimation/simulation. The collocated cokriging approach is built on a Markov-type hypothesis by which collocated secondary information is assumed to screen further away data of the same type. While Collocated Simple Cokriging with a Markov model of coregionalization is shown to perform well in estimation, an unfortunate feature of this technique when applied in simulation is variance inflation. The relatively minor variance inflation compounds in the sequential simulation process leading to a serious problem of a lack of histogram reproduction.

Chapter 5 investigates the sources of variance inflation. As a result of theoretical analysis, it is shown that the main reason for variance inflation is not the Markov assumption, but using only one collocated data in cokriging. When only one collocated data is used, the correct covariance between the new estimate and the previously calculated estimates cannot be ensured. When collocated simple cokriging is used in sequential simulation mode, the incorrect covariance between simulated data is translated into a bias of subsequent locations. This is because previously simulated data are used as conditioning data for subsequent simulated nodes. The simulated values do not have the correct covariance between each other. As a result, incorrect weights and biased estimates are obtained and variance inflation is observed.

A solution to the problem of variance inflation in collocated cokriging has been developed. Sequential simulation with cokriging based on the intrinsic model (intrinsic collocated cokriging) instead of the Markov model is considered. An intrinsic model of coregionalization is adopted with secondary data at the location being considered and at the locations of the primary data. The proposed cosimulation approach is as simple as collocated cokriging, that is, the correlation coefficient between the primary variable being modeled and secondary data is the only statistic required to integrate the secondary data in estimation/simulation. Through theoretical results and small examples it is demonstrated that the new intrinsic collocated cokriging methodology removes variance inflation, ensures reproduction of the correlation between primary and secondary data and improves the reproduction of the variogram even when the primary and secondary variables differ significantly in continuity.

To further improve practical implementation of the intrinsic collocated cokriging in the case of exhaustive multiple secondary data a novel approach to account for multiple variables simultaneously is also developed in Chapter 5. This approach is aimed at (1) merging all secondary data into a single super secondary variable, then (2) implementing intrinsic collocated cokriging with the single variable. The proposed technique results in the same results as would be obtained via "full" approach using all multiple secondary variables simultaneously. The novel super secondary methodology is very important since most commercial software allows using only one secondary data. Now users can apply super secondary simplification to obtain a geostatistical model that accounts for all multivariate secondary variables simultaneously, not only for one deemed most correlated or most relevant.

7.1.4. Multiple Univariate SGS Honoring a Correlation Matrix

Multivariate simulation is a longstanding problem in geostatistics. Fitting a model of coregionalization to many variables is difficult; however, the matrix of collocated correlation coefficients is often well informed. One option for multivariate simulation without a linear model of coregionalization is sequential Gaussian simulation with intrinsic collocated cokriging (see Chapter 5). However, this option is only applicable when the secondary data are exhaustively sampled. If this is not the case or when we are dealing with simulation of several dependent primary random variables another approach called sequential Gaussian simulation with correlated residuals is used. This simulation approach is based on performing a matrix simulation with LU decomposition of the correlation matrix at each step of sequential simulation. Modeling of each random variable is performed independently. The data on each individual random variable is used to calculate the mean and variance of the local conditional distribution for that variable using simple kriging. Simple kriging at each location is performed as many times as there are variables. The variogram models for multiple univariate simulation are modeled independently from each other. Performing a matrix simulation with LU decomposition of the correlation matrix at each step of sequential simulation (that is, multiple univariate simulation with correlated residuals) is implemented in some software. This is because multiple univariate SGS with correlated residuals has many desirable features including reproduction of the target mean and variance as well as target variogram models. The target correlation matrix, however, is not reproduced due to conditioning to local data and a combination of the variable ordering and the sequential/LU decomposition.

This thesis investigates the problem of collocated correlation being not reproduced. Chapter 6 develops a correction procedure to calculate a modified correlation matrix that leads to reproduction of the target correlation matrix. The theoretical and practical aspects of this correction are developed. The 'corrected' multiple univariate simulation with correlated residuals is shown to ensure reproduction of all target statistics, that is, mean variance and target direct variograms in several small examples and a case study. It is also noted that potential drawback of the proposed approach is possible non- positive definiteness of the correlation matrices for the residuals to be used in multiple univariate sequential Gaussian simulation to honor correlation between variables at lag 0.

7.2. Future Work

There are a number of other geostatistical problems that were not addressed in this research but may be considered in the future. The following are some ideas for future research.

7.2.1. Improving LU/P-field Simulation

In recent years a multivariate simulation approach relying on the probability field (p-field) simulation has gained attention due to its convenient implementation (see Chapter 2). P-field simulation is performed in 2 steps: (1) the local distributions of uncertainty are established at each location for each of the random variables, and (2) simulated values are drawn from these local distributions of uncertainty with correlated random numbers. The correlated random numbers are generated in such way that joint correlation matrix between random numbers corresponding to different random variables is the target matrix of correlations C. To achieve this an LU approach is frequently employed: (1) unconditional sequential Gaussian simulation is generated for each variable with respective variable variogram; (2) Cholesky decomposition of the correlation matrix C is used to obtain symmetric lower and upper triangular matrices L and U; (3) Lower triangular matrix L is multiplied by the matrix of the unconditional Sequential Gaussian realizations from (1).

P-field simulation is attractive for multivariate simulation because it separates the construction of the local distributions of uncertainty and the Monte Carlo sampling from them. However, there are two well know artifacts of p-field simulation: (1) the local conditioning data almost always appear as local minima and maxima and (2) the covariance is not reproduced in the presence of conditioning data. The reason for the covariance matrix not being reproduced is that in the presence of conditioning data, the covariance becomes non-stationary. The covariance values change depending on the closeness to conditioning data. Moreover, an LU approach itself suffers from the target correlation matrix between variables (matrix C) being not reproduced.

A new implementation of the multivariate p-field/LU approach should be proposed. A vector correlated sequential Gaussian simulation approach with non-stationary covariance should be developed to obtain the correlated random numbers to be drawn from the local distributions of uncertainty in the p-field simulation.

7.2.2. Testing for a Multivariate Gaussian Distribution

The most common simulation approach is Gaussian simulation. Each variable is transformed to a Gaussian distribution. This ensures a univariate Gaussian distribution of each variable; then, an assumption of multivariate Gaussian distribution is made. Real multivariate distributions are not likely multivariate Gaussian and show such non-Gaussian features as non-linearity and heteroscedasticity. In this case, Gaussian simulation may not reproduce important aspects of the spatial variability of the phenomenon under study. This could result in biased predictions. Therefore, it would be interesting to test how far the data depart from the multivariate Gaussian distribution; the decision of stationarity could be reconsidered or a different multivariate distribution considered. Tests for multivariate Gaussianity, however, require data independence, which is rarely the case in geostatistical modeling. Different techniques should be reviewed and a new testing methodology needs to be developed.

7.2.3. Variogram Upscaling

Integration of data from multiple sources and/or multiple scales is a common, yet challenging aspect of geostatistical modeling. Common approaches to data integration are based on a cokriging framework that often assumes the input variogram/covariance models of coregionalization are at a scale consistent with the data and the model grid. The scaling laws for the variogram have often been applied to ensure consistency of the input variogram model; however, these laws are based on a strict assumption of invariance of the variogram shape. The theoretically derived approach presented in Appendix C makes no such assumption. In fact, the examples show that there is a change in the shape of the variogram, specifically a smooth Gaussian structure at short scale can be expected with upscaling to a larger volume. This is consistent with the effects of block averaging.

A practical approach to determine directly upscaled variograms is based on a numerical integration that approximates the analytical integral of these variogram models. As with average variogram or average covariance calculations, the approximation is robust given sufficient discretization. To make this approach practically applicable a framework for direct variogram fitting should be developed. Moreover, newly developed approach presents numerous other exciting future research prospects. One area for further development is a method to downscale the block scale variogram, such that fine scale models can be constructed. This is the same objective as the work of Kupfersberger et al (1998), but the goal here will be to avoid use of the scaling laws. Another possible area of research will be to develop a linear model of coregionalization that is consistent at all scales, which could then be used to truly integrate data at different supports without any prior compositing required.

7.2.4. Accounting for the Uncertainty in Mean

An important goal of geostatistical modeling is to assess output uncertainty after processing realizations through a transfer function, in particular, to assess the uncertainty in the reserves. The decisions of stationarity and a modeling method are critical for obtaining reasonable results. Uncertainty in reserves is affected by the amount of local data and uncertainty in the modeling parameters. Oftentimes the uncertainty in the input parameters, such as mean, univariate distribution and variogram, to geostatistical model is ignored. As result, global uncertainty is underestimated. The understatement of uncertainty is especially significant for large reservoirs with sparse well control – local fluctuations above and below average cancel out and the realizations imply a very small uncertainty. Accounting for uncertainty in the parameters, especially the mean, is very important for a realistic assessment of uncertainty. There are several techniques for calculating the uncertainty in required input parameters. These include the bootstrap, and spatial bootstrap.

A standard approach to account for parameter uncertainty is to use different reference distributions in sequential simulation. Different reference distributions are obtained as direct realizations of a chosen uncertainty assessment technique. It is believed that such inclusion of parameter uncertainty and local fluctuations lead to a set of realizations that represent the local variations in reservoir parameters and larger scale uncertainty. It would be interesting to compare this commonly accepted approach with direct incorporation of the uncertainty in mean in simulation. In particular, in sequential Gaussian simulation to account for uncertainty in mean, the global mean of 0 can be modified according to its uncertainty distribution.

Bibliography

Alabert, F.G. (1987) The practice of fast conditional simulations through the LU decomposition of the covariance matrix. *Mathematical Geology*, 19:369–386.

Aki, K., and Richards, P. (1980) Quantitative seismology: W.H. Freeman and Co., San Francisco, California.

Almeida, A.S., and Journel, A.G. (1994) Joint simulation of multiple variables with a Markov-type coregionalization model. *Mathematical Geology*, 26:565–588.

Armstrong, M. (1984) Problems with universal kriging. *Journal of the International Association for Mathematical Geology*, 16: 101-108.

Armstrong, M., and Jabin, R. (1981) Variogram models must be positive-definite. *Mathematical Geology*, 13:455–459.

Boman, G., Molz, F.J., and Guven, O. (1995) An evaluation of interpolation methodologies for generating three-dimensional hydraulic property distributions from measured data. *Ground Water*, 33:247–258.

Borga, M., and Vizzaccaro, A. (1997) On the interpolation of hydrologic variables: formal equivalence of multiquadratic surface fitting and kriging. *Journal of Hydrology*, 195:160-171.

Bourennane, H., King, D., and Couturier, A. (2000) Comparison of kriging with external drift and simple linear regression for predicting soil horizon thickness with different sample densities. *Geoderma*, 97:255-271.

Bourgault, G. (1997) Using non-Gaussian distributions in geostatistical simulations. *Mathematical Geology*, 29:315-334.

Brouder, S.M., Hofmann, B.S., and Morris, D.K. (2005) Mapping soil pH: accuracy of common soil sampling strategies and estimation techniques. *Soil Science Society of America Journal*, 69: 427-442.

Brus, D.J., de Gruijter, J.J., Marsman, B.A., Visschers, R., Bregt, A.K., and Breeuwsma, A. (1996) The performance of spatial interpolation methods and choropleth maps to estimate properties at points: a soil survey case study. *Environmetrics*, 7:1–16.

Caers, J. (2000) Adding local accuracy to direct sequential simulation. *Mathematical Geology*, 32:815-850.

Chiles, J.P., and Delfiner, P. (1999) Geostatistics: modeling spatial uncertainty. Wiley, New York.

Chu, J.C. (1996) Fast sequential indicator simulation: Beyond reproduction of indicator variograms. *Mathematical Geology*, 28:923-936.

Cressie, N.A. (1993) Statistics for spatial data. John Wiley & Sons, Ontario, Canada.

Cressie, N.A., and Hawkins, D.M. (1980) Robust estimation of the variogram.1. *Journal of the International Association for Mathematical Geology*, 12:115–125.

Cockx, L., Van Meirvenne, M., and De Vos, B. (2007) Using the EM38DD soil sensor to delineate clay lenses in a sandy forest soil. *Soil Science Society of America Journal*, 71:1314-1322.

Caruso, C., and Quarta, F. (1998) Interpolation methods comparison. *Computers & Mathematics with Applications*, 35:109–126.

Creutin, J.D., and Obled, C. (1982) Objective analyses and mapping techniques for rainfall fields: an objective comparison. *Water Resources Research*, 18:413–431.

David, M. (1977) Geostatistical ore reserve estimation. Elsevier, Amsterdam.

Davis, M.W. (1987) Production of conditional simulations via the LU decomposition of the covariance matrix. *Mathematical Geology*, 19:91–98.

Declercq, F.A.N. (1996) Interpolation methods for scattered sample data: accuracy, spatial patterns, processing time. *Cartography and Geographic Information Systems*, 23:128–144.

Deutsch, C.V. (2006) What is the reservoir geostatistics good for?. *Journal of Canadian Petroleum Technology*, 45:14-20.

Deutsch, C.V. (2002) Geostatistical reservoir modeling. Oxford University Press, New York.

Deutsch, C.V., and Journel, A.G. (1998) GSLIB: geostatistical software library and users guide. Oxford University Press, New York.

Deutsch, C.V. (1996) Direct Assessment of Local Accuracy and Precision, Geostatistics Wollongong '96, 1, Baafi and Schofield, editors, Kluwer Academic Publishers: 115-125.

Deutsch, C.V. (1994) Kriging with strings of data. *Mathematical Geology*, 26:623-638.

Deutsch, C.V. (1993) Kriging in a finite domain. *Mathematical Geology*, 25:41-52.

Dingman, S.L. (1994) Physical hydrology. Macmillan College Publishing Company.

Diodato, N., and Ceccarelli, M. (2005) Interpolation processes using multivariate geostatistics for mapping of climatological precipitation mean in the Sannio Mountains (southern Italy). *Earth Surface Processes and Landforms*, 30:259-268.

Dirks, K.N., Hay, J.E., Stow, C.D., and Harris, D. (1998) High-resolution studies of rainfall on Norfolk Island, part II: interpolation of rainfall data. *Journal of Hydrology*, 208:187–193.

Dowd, P.A. (1982) Lognormal kriging - the general case. *Mathematical Geology*, 14:475–498.

Doyen, P.M., den Boer, L.D., and Pillet, W.R. (1996) Seismic porosity mapping in the Ekofisk Field using a new form of collocated kriging. SPE 36498.

Falivene, O., Cabrera, L., and Saez, A. (2007) Optimum and robust 3D facies interpolation strategies in a heterogeneous coal zone (Tertiary As Pontes Basin, NW Spain). *International Journal of Coal Geology*, 71:185–208.

Franke, R. (1982) Scattered data interpolation: tests of some methods. *Mathematics of Computation*, 38:181-200.

Frykman, P., and Deutsch, C.V. (2002) Practical application of geostatistical scaling laws for data integration. *Petrophysics*, 43:153-171

Frykman, P., and Deutsch, C.V. (1999) Geostatistical scaling laws applied to core and log data. SPE 56822.

Emery, X. (2006) Multigaussian kriging for point-support estimation: incorporating constraints on the sum of the kriging weights. *Stochastic Environmental Research and Risk Assessment*, 20: 53-65.

Emery, X. (2005) Simple and ordinary Multigaussian kriging for estimating recoverable reserves. *Mathematical Geology*, 37:295-319.

Emery, X. (2004) Properties and limitations of sequential indicator simulation. *Stochastic Environmental Research and Risk Assessment*, 18:414-424.

Galli A., and Meunier G. (1987) Study of a gas reservoir using the external drift method. Geostatistical case studies. G. Matheron and M. Armstrong (eds), 105-120.

Gallichand, J., and Marcotte, D. (1993) Mapping clay content for subsurface drainage in the Nile Delta. *Geoderma*, 58:165–179.

Goulard, M., and Voltz, M. (1992) Linear coregionalization model – tools for estimation and choice of cross-variogram matrix. *Mathematical Geology*, 24:269-286.

Grimm, J.W., and Lynch, J.A. (1991) Statistical analysis of errors in estimating wet deposition using five surface estimation algorithms. *Atmospheric Environment*, 25:317–327.

Gringarten, E., and Deutsch, C.V. (2001) Teacher's aide - variogram interpretation and modeling. *Mathematical Geology*, 33:507–534.

Goovaerts, P. (2002) Geostatistical modelling of spatial uncertainty using p-field simulation with conditional probability fields. *International Journal of Geographical Information Science*, 16:167-178.

Goovaerts, P. (2000): Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of Hydrology*, 228:113–129.

Goovaerts, P. (1997) Geostatistics for natural resources evaluation. Oxford University Press, New York.

Gotway, C.A., Ferguson, R.B., Hergert, G.W., and Peterson, T.A. (1996) Comparison of kriging and inverse-distance methods for mapping soil parameters. *Soil Science Society of America Journal*, 60:1237–1247.

Isaaks, E.H. (1990) The application of Monte Carlo methods to the analysis of spatially correlated data. PhD thesis, Standford University, Standford, CA.

Isaaks, E.H., and Srivastava, R.M. (1989) An introduction to applied aeostatistics. Oxford University Press, New York.

Jensen, O.P., and Miller, T.J. (2005) Geostatistical analysis of the abundance and winter distribution patterns of the blue crab Callinectes sapidus in Chesapeake Bay. *Transaction of the American Fisheries Society*, 134: 1582-1598.

Journel, A.G., and Kyriakidis, P.C. (2004) Evaluation of mineral reserves: a simulation approach. Oxford University Press, New York

Journel, A.G., Kyriakidis, P.C., and Mao, S. (2000) Correcting the smoothing effect of estimators: a spectral postprocessor. *Mathematical Geology*, 32:787–813.

Journel, A.G. (1999) Markov models for cross-covariances. *Mathematical Geology*, 31:955-964.

Journel, A.G. (1990) Geostatistics for reservoir characterization. SPE 20750.

Journel, A.G. (1986) Geostatistics - models and tools for the earth sciences. *Mathematical Geology*, 18:119–140.

Journel, A.G., and Isaaks, E.H. (1984) Conditional indicator simulation: application to a Saskatchewan uranium deposit. *Mathematical Geology*, 16:685–718.

Journel, A.G. (1980) The lognormal approach to predicting local distributions of selective mining unit grades. *Mathematical Geology*, 12:285–303.

Journel, A.G., and Huijbrogts, C.J. (1978) Mining geostatistics. Academic Press, London.

Journel, A.G. (1974) Geostatistics for conditional simulation of ore bodies. *Economic Geology*, 69:673-687.

Haas, A., Biver, P., and Mouliere, D. (1998) Simulations stochastiques en cascade (Stochastic sequential simulation). Cahiers de Geostatistique 6, Ecole des Mines de Paris, 31–43.

Hosseini, E., Gallichand, J., Marcotte, D. (1994) Theoretical and experimental performance of spatial interpolation methods for soil salinity analysis. *Transactions of the ASAE*, 37:1799–1807.

Huijbregts, C.J., and Matheron, G. (1971) Universal kriging - an optimal approach to trend surface analysis. In Decision Making in the Mineral Industry, 159–169. Canadian Institute of Mining and Metallurgy, Special Volume 12.

Hodgson, M.E. (1992) Sensitivity of spatial interpolation models to parameter variation: ACSM Technical Papers—Albuquerque, American Congress on Surveying and and Mapping, Bethesda, MD, 2:113–122.

Hutchinson, M.F. (1993) On thin plate splines and kriging. In: Tarter ME, Lock MD (eds) Computing and science in statistics. University of California, Berkeley, CA.

Kishne, A.S., Bringmark, E., Bringmark, L., and Alriksson, A. (2003) Comparison of ordinary and lognormal kriging on skewed data of total cadmium in forest soils of Sweden. *Environmental Monitoring and Assessment*, 84:243-263.

Kravchenko, A.N. (2003) Influence of spatial structure on accuracy of interpolation methods. *Soil Science Society of America Journal*, 67:1564-1571.

Kravchenko, A.N., Boast, C.W., and Bullock, D.G. (1999) Multifractal analysis of soil spatial variability. *Agronomy Journal*, 91:1033-1041.

Kupfersberger, H., Deutsch, C.V., and Journel, A.G. (1998) Deriving constraints on small-scale variograms due to variograms of large-scale data. *Mathematical Geology*, 30:837-852.

Kyriakidis, P.C., and Journel, A.G. (2001) Stochastic modeling of atmospheric pollution: a spatial time-series framework. Part I: methodology. *Atmospheric Environment*, 35:2331-2337.

Lark, R.M., and Papritz, A. (2003) Fitting a linear model of coregionalization for soil properties using simulated annealing. *Geoderma*, 115:245-260.

Laslett, G.M. (1994) Kriging and splines: an empirical comparison of their predictive performance in some applications. *Journal of the American Statistical Association*, 89:391–409.

Laslett, G.M., and McBratney, A.B. (1990) Further comparison of spatial methods for predicting soil pH. *Soil Science Society of America Journal*, 54:1553–1558.

MacDougall, E.B. (1976) Computer programming for spatial problems. John Wiley & Sons, New York.

Mardikis, M.G., Kalivas, D.P., and Kollias, V.J. (2005) Comparison of interpolation methods for the prediction of reference evapotranspiration - an application in Greece. *Water Resources Management*, 19:251–278.

Marechal, A. (1984) Kriging seismic data in presence of faults. In Verly G et al (eds) Geostatistics for natural resources characterization. Reidel, Dordrecht, Holland.

Martinez-Cob, A. (1996) Multivariate geostatistical analysis of evapotranspiration and precipitation in Mountain Terrain. *Journal of Hydrology*, 174:19–35.

Moinard, L. (1987) Application of kriging to the mapping of a reef..... Geostatistical case studies. G. Matheron and M. Armstrong (eds), 93-104.

Morrison, J.L. (1971) Method-produced error in isarithmic mapping. American Congress on Surveying and Mapping, Washington, DC, 76 p.

Moyeed, R.A., and Papritz, A. (2002) An empirical comparison of kriging methods for nonlinear spatial point prediction. *Mathematical Geology*, 34:365-386.

Mueller, T.G., Dhanikondam, S.R.K., Pusuluri, N.B., Karathanasis, A.D., Mathias, K.K., Mijatovic, B., and Sears, B.G. (2005) Optimizing inverse distance weighted interpolation with cross-validation. *Soil Science*, 170:504-515.

Mueller, T.G., Pusuluri, N.B., Mathias, K.K., Cornelius P.L., Barnhisel, R.I., and Shearer, S.A. (2004) Map quality for ordinary kriging and inverse distance weighted interpolation. *Soil Science Society of America Journal*, 68:2042-2047.

Myers, D.E. (1983) Estimation of linear combinations and co-kriging. *Mathematical Geology*, 15:633-637.

Myers, D.E. (1982) Matrix formulation of co-kriging. *Mathematical Geology*, 14:249-257.

Nader, I.A., and Wein, R.W. (1998) Spatial interpolation of climatic normals: test of a new method in the Canadian Boreal Forest. *Agricultural and Forest Meteorology*, 92:211-225.

Olea, R.A. (1973) Optimal countour mapping using universal kriging. *Transactions - American Geophysical Union*, 54:1288-1288.

Oz, B., Deutsch, C.V., Tran, T.T., and Xie, Y. (2003) DSSIM-HR: A Fortran 90 program for direct sequential simulation with histogram reproduction. *Computers & Geosciences*, 29:39-51.

Pelletier, B., Dutilleul, P., Larocque, G., and Fyles, J.W. (2004) Fitting the linear model of coregionalization by generalized least squares. *Mathematical Geology*, 36:323-343.

Peucker, T.K. (1980) The impact of different mathematical approaches to contouring. *Cartographica*, 17: 73–95.

Phillips, D.L., Lee, E.H., Herstrom, A.A., Hogsett, W.E., and Tingey, D.T. (1997) Use of auxiliary data for spatial interpolation of ozone exposure in southeastern forests. *Environmetrics*, 8:43–61.

Pople, A.R., Phinn, S.R., Menke, N., Grigg, G.C., Possingham, H.P., and McAlpine, C. (2007) Spatial patterns of kangaroo density across the South Australian pastoral zone over 26 years: aggregation during drought and suggestions of long distance movement. *Journal of Applied Ecology*, 44:1068-1079.

Rana, G., and Katerji, N. (2000) Measurement and estimation of actual evapotranspiration in the field under Mediterranean climate: a review. *European Journal of Agronomy*, 13:125–153.

Ren, W, Leuangthong, O, and Deutsch, C.V. (2007) Global resource uncertainty using a spatial/multivariate decomposition approach. *Journal of Canadian Petroleum Technology*, 46:33-39.

Rivoirard, J. (2002) On the structural link between variables in kriging with external drift. *Mathematical Geology*, 34:797-808.

Rivoirard, J., (2001) Which models for collocated cokriging?. *Mathematical Geology*, 33:117-131.

Roth, C. (1998) Is lognormal kriging suitable for local estimation?. *Mathematical Geology*, 30:999-1009.

Rojas-Avellaneda, D., and Silvan-Cardenas, J.L. (2006) Performance of geostatistical interpolation methods for modeling sampled data with non-stationary mean. *Stochastic Environmental Research and Risk Assessment*, 20:455-467.

Rouhani, S. (1986) Comparative study of ground-water mapping technique. *Ground Water*, 24:207–216.

Saito, H., McKenna, S.A., Zimmerman, D.A., and Coburn, T.C. (2005) Geostatistical interpolation of object counts collected from multiple strip transects: ordinary kriging versus finite domain kriging. *Stochastic Environmental Research and Risk Assessment*, 19:71-85.

Saito, H., and Goovaerts, P. (2002) Accounting for measurement error in uncertainty modeling and decision-making using indicator kriging and p-field simulation: application to a dioxin contaminated site. *Environmetrics* 13:555-567.

Shepard, D. (1968) A Two-dimensional interpolation function for irregularly spaced data. Proceedings of the 1968 23rd ACM, ACM Press, New York.

Schloeder, C.A., Zimmerman, N.E., and Jacobs, M.J. (2001) Comparison of methods for interpolating soil properties using limited data. *Soil Science Society of America Journal*, 65:470–479.

Seguret, S. (1989) Filtering periodic noise by using trigonometric kriging. Geostatistics, Armstrong (ed.),. 481-492.

Shmaryan, L.E., and Journel, A.G. (1999) Two Markov models and their application. *Mathematical Geology*, 31:965–988.

Soares, A. (2001) Direct sequential simulation and cosimulation. *Mathematical Geology*, 33:911-926.

Tabios, G.Q., and Salas, J.D. (1985) A comparative analysis of techniques for spatial interpolation of pecipitation. *Water Resources Bulletin*, 21:365–380.

Teegavarapu, R.S.V., and Chandramouli, V. (2005) Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. *Journal of Hydrology*, 312:191–206.

Vauclin, M., Vieira, S.R., Vachaud, G., Nielsen, D.R. (1983) The use of cokriging with limited field soil observations. *Soil Science Society of America Journal*, 47:175-184.

Verly, G. (1983) The Multigaussian approach and its applications to the estimation of local reserves. *Mathematical Geology*, 15:259–286.

Voltz, M., and Webster, R. (1990) A comparison of kriging, cubic splines and classification for predicting soil properties from sample information. *Journal of Soil Science*, 41:473–490.

Wackernagel, H. (2003): Multivariate geostatistics. Springer, Berlin New York.

Wackernagel, H. (1994) Cokriging versus kriging in regionalized multivariate data nalysis. *Geoderma*, 62:83–92.

Wahba, G. (1990) Spline models for observational data. In: CBMS-NSF Regional conference series in applied mathematics. Society for Industrial and Applied Mathematics, Philadelphia, PA, 169 p.

Weber, D.D., and Englund, E.J. (1994) Evaluation and comparison of spatial interpolators II. *Mathematical Geology*, 26:589-603.

Weber, D.D., and Englund, E.J. (1992) Evaluation and comparison of spatial interpolators. *Mathematical Geology*, 24:381-391.

Weisz, R., Fleischer, S., and Smilowitz, Z. (1995) Map generation in high-value horticultural integrated pest management: appropriate interpolation methods for site-specific pest management of Colorado Potato Beetle (Coleoptera: Chrysomelidae). *Journal of Economic Entomology*, 88:1650–1657.

Wollenhaupt, N.C., Wolkowski, R.P., and Clayton, M.K (1994) Mapping soil test phosphorus and potassium for variable-rate fertilizer application. *Journal of Production Agriculture*, 7:441–448.

Xu, W., Tran, T.T., Srivastava, R.M., and Journel, A.G. (1992) Integrating seismic data in reservoir modeling: the collocated cokriging alternative. SPE 24742.

Yamamoto, J.K. (2005) Correcting the smoothing effect of ordinary kriging estimates. *Mathematical Geology*, 37:69-94.

Zimmerman, D., Pavlik, C., Ruggles, A., and Armstrong, M.P. (1999) An experimental comparison of ordinary and universal kriging and inverse distance weighting. *Mathematical Geology*, 31:375-390.

APPENDIX A

Statistical Approach to Inverse Distance Interpolation

Inverse distance interpolation is a robust and widely used estimation technique. Variants of kriging are often proposed as statistical techniques with superior mathematical properties such as minimum error variance; however, the robustness and simplicity of inverse distance interpolation motivate its continued use. This Appendix A presents an approach to integrate statistical controls such as minimum error variance into inverse distance interpolation. The optimal exponent and number of data may be calculated globally or locally. Measures of uncertainty and local smoothness may be derived from inverse distance estimates.

A.1. Introduction

One of the most important problems in the geo- and environmental sciences is spatial prediction. Spatial predictions are required for planning, risk assessment, and decision-making. Typical applications include determining the profitability of mining an orebody, management of soil resources, pest management, designing a network of environmental monitoring stations, and quantifying the uncertainties inherent in spatial predictions (Weisz et al, 1995; Gotway et al., 1996; Moyeed and Papritz, 2002).

Spatial prediction techniques, also known as spatial interpolation techniques, differ from classical modeling approaches in that they incorporate information on the geographic position of the sample data points (Journel and Huijbregts, 1978; Cressie, 1993). Spatial predictions offer means of describing a variety of responses over different spatial scales (Schloeder, et al., 2001). They provide a unique and smooth property distribution that reproduces the sample points (conditioning data); spatial prediction techniques aim at the local accuracy of resulting uncertainty distributions (Isaaks and Srivastava, 1989; Journel et al., 2000). The most common interpolation techniques calculate the estimates for a property at any given location by a weighted average of nearby data. Weighting is assigned either according to deterministic or statistical (spatial covariance) criteria. When a statistical criterion is used, the field is considered as a random process and the optimality of the averaging method is determined in terms of minimizing the estimation variance. When a deterministic criterion is used, the measures of optimality are arbitrarily chosen (Borga and Vizzaccaro, 1997). Among statistical methods, geostatistical kriging-based techniques, including Simple and Ordinary Kriging, Universal Kriging and Simple Cokriging (see Journel, 1986; Cressie, 1993; Deutsch, 2002) have been often used for spatial analysis. Among deterministic methods, Inverse Distance Weighted interpolation and its modifications (see Franke, 1982; Nader and Wein, 1998) are the most often applied.

In this Appendix A, we expand the applicability of inverse distance methods by introducing a statistical formalism. The proposed formalism is based on the assumption of stationarity and is aimed at providing the estimation variance at the unsampled locations as a measure of accuracy. We propose a general approach to find the optimal exponent value and the optimal number of neighboring points to be used in estimation. Inverse distance interpolation is very sensitive to the number of data used in interpolation and to the exponent value; a significant improvement in estimation precision can be achieved by selecting optimal values (Kravchenko et al., 1999). Presently, however, there is no exact recommendation about the choice of exponent value and the optimal number of neighboring points to be used in the inverse distance estimation. A number of researchers approached this problem and their recommendations are contradictory. For example, in the case of the inverse distance squared interpolation Morrison (1974), MacDougall (1976), Peucker (1980), and Hodgson (1992) recommended to use respectively, $3 \leq k \leq 7$, $6 \leq k \leq 9$, $k \leq 6$, and $4 \leq k \leq 7$ data; however, Declercq (1996), recommended $4 \leq k \leq 8$ for "smooth" surfaces and $16 \leq k \leq 24$ for abruptly changing surfaces. There exists also an approach for optimizing inverse distance weighted interpolation globally by selecting distance exponent values that minimize cross-validation (or jackknife) errors of prediction (Muleller et al., 2005; Rojas-Avellaneda and Silvan-Cardenas, 2006). The effectiveness of this approach has not been critically evaluated.

In this Appendix A we document the sensitivity of the inverse distance estimation to the number of data and the exponent used in estimation. As direct result of the sensitivity analysis, a local inverse distance interpolation approach is proposed to create estimates with minimum achievable estimation variance for the inverse distance interpolation.

A.2. Background: Kriging versus Inverse Distance Interpolation

Kriging is a well-proven technique that provides the best linear unbiased estimate and its variance at the unknown location. It is an exact interpolator in the sense that the estimation at a data location returns the original data value. In theory, kriging is a statistically optimal interpolator in the sense that it minimizes estimation variance when the variogram (measure of spatial continuity of the variable under study) is known and under the assumption of stationarity.

Inverse distance weighting estimates the variable of interest by assigning more weight to closer points. It is a simple technique that does not require prior information, that is, variogram model, to be applied to spatial prediction. Despite this simplicity, inverse-distance estimators are shown (experimentally) to be quite sensitive to the type of database or data characteristics (e.g., skewed/unskewed distribution, isotropic/anisotropic phenomenon, on regular grid/clustered data, etc.), to the number of neighbors used in the estimate, and to the exponent of distance used in weighting (Weber and Englund, 1994). In practical applications, inverse distance weighted interpolation may be preferred over kriging-based techniques when there is a problem of making meaningful estimates of the field spatial structure from sparse data (Wahba, 1990; Hutchinson, 1993). It is also used when a quick visualisation of the variable under study is required (Borga and Vizzaccaro, 1997). Moreover, a large number of comparative studies among different interpolators found that depending upon the situation at hand, inverse distance weighting can be as good or better than geostatistical kriging-based techniques (Weber and Englund, 1992; Gallichand and Marcotte, 1993; Dingman, 1994; Boman et al., 1995; Brus et. al., 1996; Declercq, 1996; Dirks et al., 1998; Moyeed and Papritz, 2002; Kravchenko, 2003; Mueller et al., 2004; Brouder et al., 2005). These studies were based on geologically sound visual appearance; cross validation and jackknife, which involves consecutively removing a data value from the sample data set and interpolating to that site using the remaining conditioning data values, then comparing the estimated values against the true data (Isaaks and Srivastava, 1989); robustness; or measures of response variables derived from the interpolated property. Therefore, it may be important to analyze the inverse distance weighted interpolation approach in greater detail with the aim of improving it.

The main advantages of kriging over inverse distance interpolation are cited as (1) robustness of estimates with respect to the number of data used in estimation (four nearest neighbors appear to be generally inadequate for kriging; however, change in the kriging estimates between 12 and 20 neighbors is minimal), (2) ability to take into account the spatial structure of the data points (anisotropy) and (3) availability of the estimation variance that yields a measure of the accuracy of any single interpolated value. This measure can have a dual role. Firstly, it evaluates the reliability of our estimates. Secondly, it can serve as a guideline to identify the most uncertain areas for further measurements (Rouhani, 1985).

A.3. Inverse Distance Interpolation

An inverse distance interpolation is one of the simplest and most popular interpolation techniques. It combines the proximity concept with the gradual change of the trend surface. An inverse distance (ID) weighted interpolation is defined as a spatially weighted average of the sample values within a search neighborhood (Shepard, 1968; Franke, 1982; Diodato and Ceccarelli, 2005). It is calculated as

$$Z^*(\mathbf{u}) = \sum_{i=1}^{n} \lambda_i Z(\mathbf{u}_i), \tag{A.1}$$

where $\mathbf{u}$ is the estimation location, $\mathbf{u}_i, i = 1,\ldots,n,$ are the locations of the sample points within the search neighborhood, $Z^*(\mathbf{u})$ is the inverse distance estimate at the estimation location, n is the number of sample points, $\lambda_i, i = 1,\ldots,n,$ are the weights assigned to each sample point, and $Z(\mathbf{u}_i), i = 1,\ldots,n,$ are the conditioning data at sample points. The weights are determined as

$$\lambda_i = \frac{\left(\frac{1}{d_i^p}\right)}{\sum_{i=1}^{n}\left(\frac{1}{d_i^p}\right)}, \quad (i = 1,\ldots,n), \tag{A.2}$$

where d_i are the Euclidian distances between estimation location and sample points, and exponent p is the power or distance exponent value. Note that the sum of the inverse distance weights $\lambda_i, i = 1,\ldots,n,$ is equal to 1, that is,

$$\sum_{i=1}^{n} \lambda_i = 1.$$

The most common value applied for the power p is 2; then estimator in (A.1)-(A.2) is called inverse squared distance (ISD) interpolator. However, any value for p can be chosen. As p increases, the interpolated value by inverse distance is assigned the value of the nearest sample point, that is, inverse distance estimate becomes the same as estimate produced by polygonal method. (Diadato and Ceccarelli, 2005). Several modifications of the inverse distance method are developed including gradient inverse distance interpolation (GIDW) (Nalder and Wein, 2000).

7

A.4. Statistical Formalism

A.4.1 Stationarity

The uncertainty in the true value at an usampled location $z(\mathbf{u}) \in A$ can be modeled using cumulative probability distribution function of a random variable $Z(\mathbf{u})$,

$$F(\mathbf{u}; z) = \mathrm{Prob}\{Z(\mathbf{u}) \leq z\}.$$

This probability distribution function can be thought of as being a model of the lack of knowledge about the value of the variable under study at the usampled location **u**. Repetitive samples are needed to infer any statistic. Unfortunately, in the spatial context repetitive samples are not available. A measurement cannot be repeated at the same location **u** to obtain probability distribution of the random variable $Z(\mathbf{u})$. Stationarity is a two part decision required for statistical prediction: (1) a decision to pool data for common analysis, and (2) a decision of the location-independence of the random function probability distribution and all its moments by translation over the domain A. First order of stationarity assumes that the mean of the variable of interest is constant throughout the domain A; second order of stationarity assumes that the variance of data is constant throughout the study domain A (Deutsch, 2002). That is,

$$\begin{aligned} \mathrm{E}(Z(\mathbf{u})) &= m, \quad \forall \mathbf{u} \in A; \\ \mathrm{Var}(Z(\mathbf{u})) = \mathrm{E}(Z(\mathbf{u}) - m(\mathbf{u}))^2 &= \sigma^2, \quad \forall \mathbf{u} \in A. \end{aligned} \tag{A.3}$$

4.2. Mean and Variance of the Inverse Distance Estimator for the Stationary Domain

The mean and variance of the inverse distance estimator $Z^*(\mathbf{u})$ at estimation location **u** given by (A.1)-(A.2) can be derived under the assumption of stationarity as follows

$$\mathrm{E}(Z^*(\mathbf{u})) = \mathrm{E}\left(\sum_{i=1}^{n}\lambda_i Z(\mathbf{u}_i)\right) = \sum_{i=1}^{n}\lambda_i \mathrm{E}(Z(\mathbf{u}_i)) = m\sum_{i=1}^{n}\lambda_i = m;$$

$$\mathrm{Var}(Z^*(\mathbf{u})) = \mathrm{Var}\left(\sum_{i=1}^{n}\lambda_i Z(\mathbf{u}_i)\right) = \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j \mathrm{Cov}(Z(\mathbf{u}_i), Z(\mathbf{u}_j)); \tag{A.4}$$

where $\mathrm{Cov}(Z(\mathbf{u}_i), Z(\mathbf{u}_j))$, $i, j = 1, \cdots, n$, denotes data-to-data covariance function calculated under assumption of stationarity though the semivariogram model $2\gamma(\mathrm{h})$ (Journel and Huijbregts, 1978).

The estimate and variance of the inverse distance estimator at the data location are set to the data value at that location and stationary domain variance σ^2, respectively. Note, however, that despite neither the IDW estimate and variance at the data location are defined; it can be shown that they converge in a limit to the data value at that precise location and stationary domain variance σ^2, respectively.

Under stationarity, the variance at each location of the domain should be exactly equal to the stationary domain variance σ^2. However, the map of the inverse distance estimates is smooth. The smoothing effect of inverse distance interpolation technique is directly related to the IDW variance via this expression

$$\text{Smoothing effect} = \sigma^2 - \mathrm{Var}(Z^*(\mathbf{u})) = \sigma^2 - \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j \mathrm{Cov}(Z(\mathbf{u}_i), Z(\mathbf{u}_j)). \tag{A.5}$$

Note that smoothing effect of the inverse distance interpolator in (A.5) can be also referred to as the *missing variance*. This is because by adding the variable with variance (A.5) to the inverse distance estimate we will obtain new variable with variance equal to the stationary domain variance σ^2. Note that at the data locations missing variance is equal to zero.

Moreover, note that the estimation (error) variance at the estimation location **u** for the inverse distance estimator can be calculated (under the assumption of stationarity) as (Deutsch, 2002)

$$\sigma_{est}^2 = \mathrm{E}[Z - Z^*(\mathbf{u})]^2 = \sigma^2 - 2\sum_{i=1}^{n}\lambda_i \mathrm{Cov}(Z(\mathbf{u}_j), Z(\mathbf{u})) + \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j \mathrm{Cov}(Z(\mathbf{u}_i), Z(\mathbf{u}_j)). \tag{A.6}$$

A.5. Sensitivity of the Inverse Distance Weighted Interpolation to the Number of Data: Example

A.5.1. Data

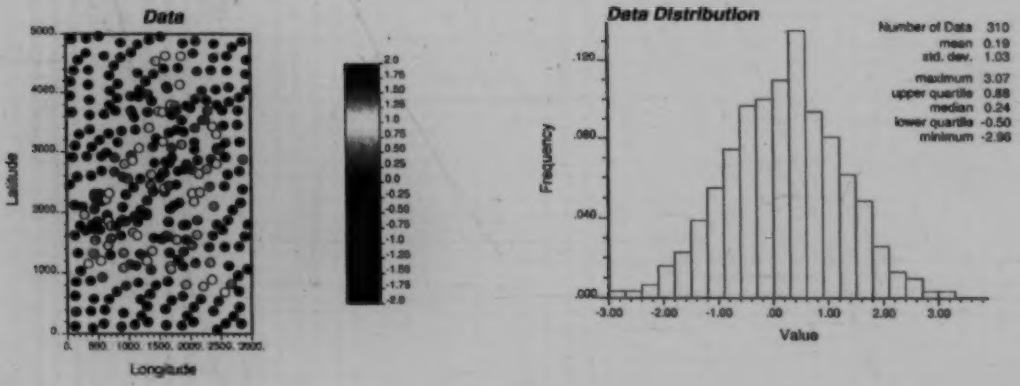
There are a total of 310 samples within a 2D rectangular project area extending 3km in the Easting X direction (longitude) and 5km in the Northing Y direction (latitude). The samples are in normal score units. The location map of the Gaussian values together with their histogram is given in Figure A.1. The experimental omnidirectional variogram of the data together with its fit (isotropic spherical with nugget effect of zero and range of correlation 1450 meters) is shown in Figure A.2. To calculate the experimental variogram the following parameters were used: number of lags was set to 12; lag separation distance and lag tolerance were set to 200m and 100m, respectively; horizontal bandwidth was set to 5000m.

A.5.2. Estimation

Figure A.3 shows results of the inverse distance interpolation for the mean and variance (that is, missing variance given in (A.5)) of the local conditional distributions obtained based on 3 data with several exponent values, that is, $p = 1,2,3,9$. Figure A.4 shows analogous results of the inverse distance interpolation but obtained based on 24 data with different exponent values.

It can be clearly noted from Figures A.3-A.4 that

- With increase in the number of data and smoothness of the map, the variance of the local conditional distributions increases;
- With increase in the power exponent the variance of the local conditional distributions decreases, the variance of the estimated values approaches stationary domain variance everywhere except at the boundaries between higher/lower values.



a) b)

Figure A.1: Location map of 310 samples (a) together with their distribution (b).

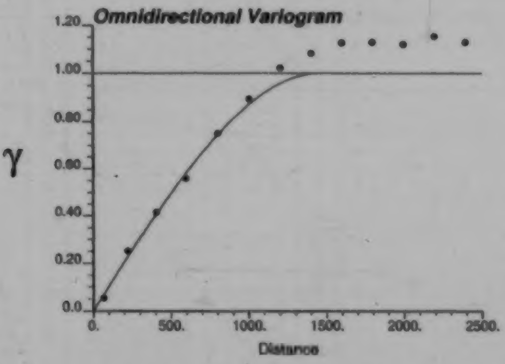
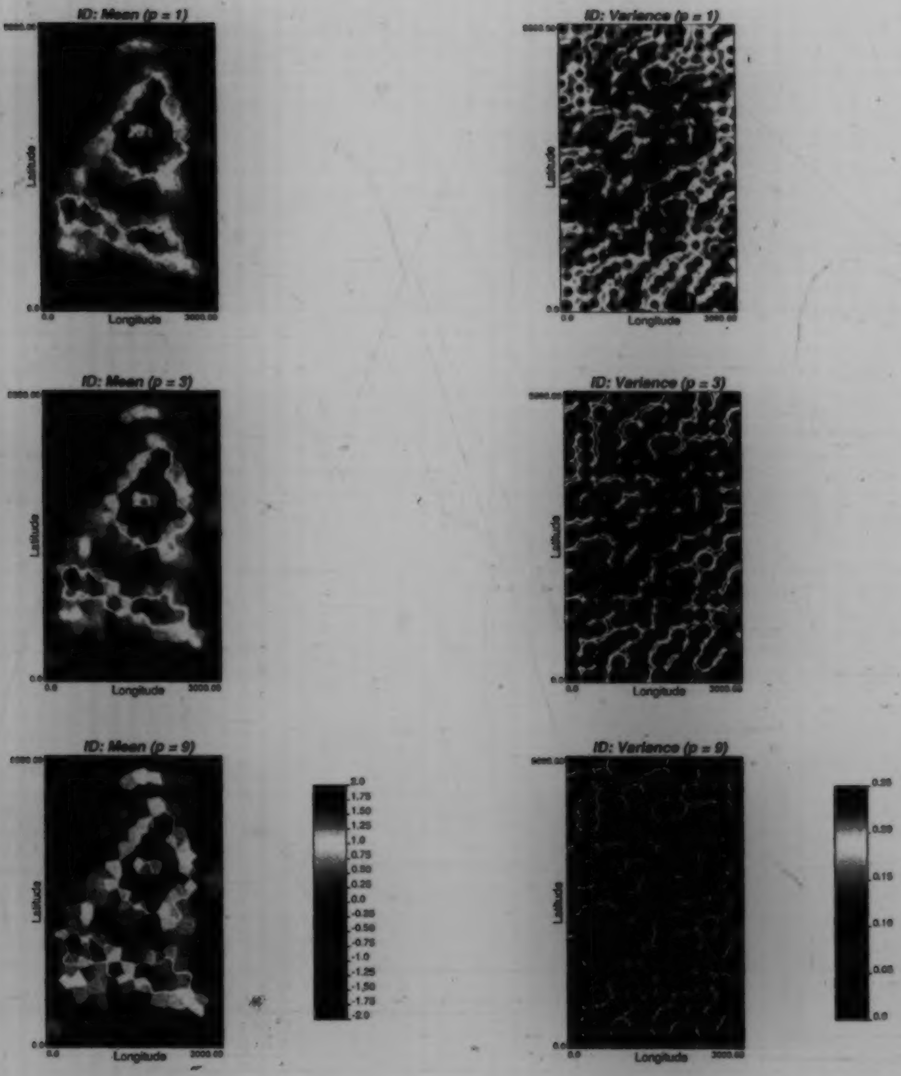
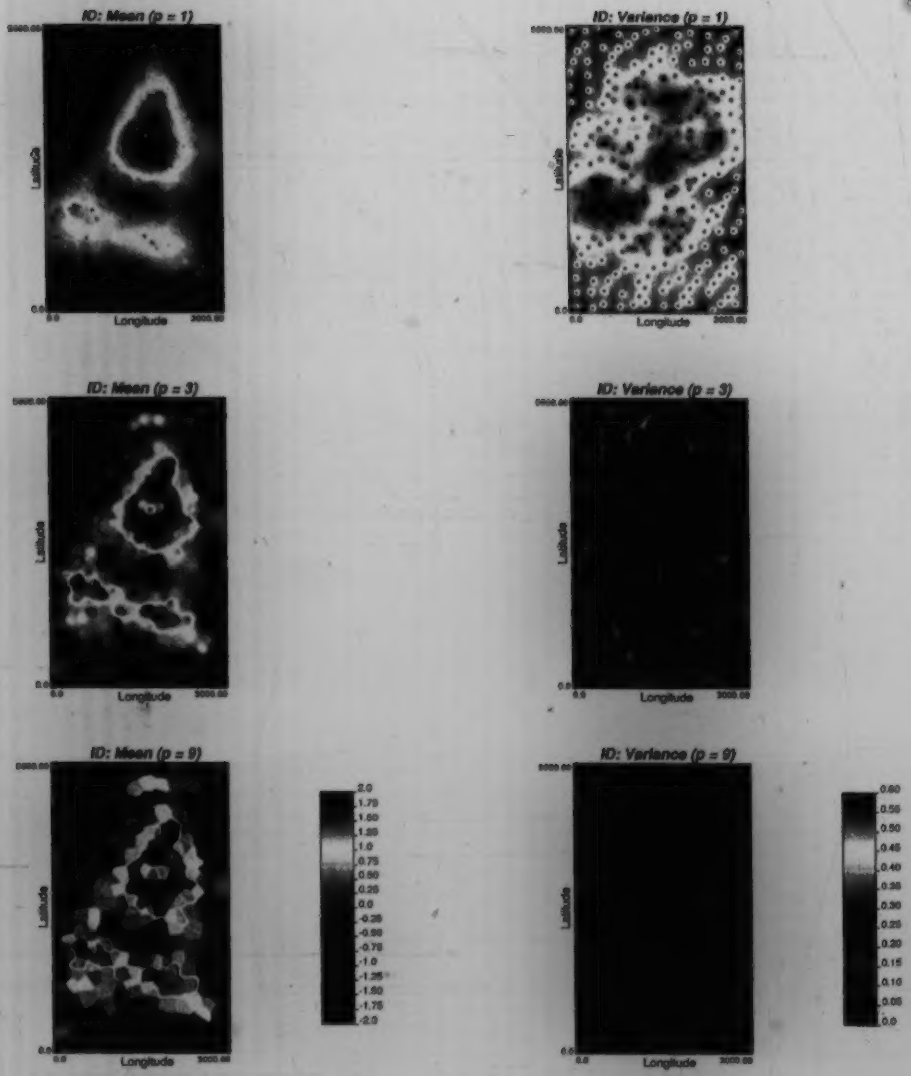


Figure A.2: Experimental omnidirectional variogram (points) together with its variogram fit for 310 samples in the study domain.


ID: Mean (p = 1)
ID: Variance (p = 1)
ID: Mean (p = 3)
ID: Variance (p = 3)
ID: Mean (p = 9)
ID: Variance (p = 9)
Latitude
Longitude
0.0
3000.00
2.0
1.75
1.50
1.25
1.0
0.75
0.50
0.25
0.0
-0.25
-0.50
-0.75
-1.0
-1.25
-1.50
-1.75
-2.0
0.25
0.20
0.15
0.10
0.05
0.0

a) b)

Figure A.3: Results of the inverse distance interpolation for the mean (a) and variance (b) of the local conditional distributions obtained based on 3 data with exponent value *p* equal to: 1 (top); 3 (middle) and 9 (bottom).



a) b)

Figure A.4: Results of the inverse distance interpolation for the mean (a) and variance (b) of the local conditional distributions obtained based on 24 data with exponent value p equal to: 1 (top); 3 (middle) and 9 (bottom).

A.5.3. Sensitivity

To analyze results of estimation in greater detail a slice at X = 100 is selected. The estimation variances for the inverse distance interpolator with exponent value of 1 (p = 1) as a functions of the number of data for the chosen slice is shown in Figure A.5.

Looking at Figure A.5 we can clearly note that with increase in the number of data for the inverse distance interpolator there is generally an increase in the estimation variance. Note that the increase in the inverse distance estimation variance is quite substantial when using 24 data instead of 3. Looking at Figure A.5 we can also conclude that for the slice at slice at X = 100 in order to minimize the estimation variance, a small number of values (3 data) should be used for estimation.

The estimation variances for the inverse distance interpolator obtained based on 3 and 24 data as functions of the exponent value for the slice at X = 100 are shown in Figure A.6. Note that when smaller number of data is used for interpolation, the difference in the estimation variance is minimal. Large power exponents (p = 9) produce estimates with larger estimation variance.

On the other hand, when larger number of data is used for interpolation, the difference in the estimation variance for different exponent values is more pronounced. The inverse distance interpolation with higher exponent value (p = 9) is producing better result.

A.6. Inverse Distance with Locally Varying Parameters: Small Example

The optimal inverse distance weighted interpolation parameters, that is, number of data and power exponent, can be chosen by minimizing estimation variance at each location. Depending on the estimation location, of course, optimal parameters will be different as well as will be different estimation variance. Estimation variance obtained is the minimum estimation variance achievable by the inverse distance interpolation technique.

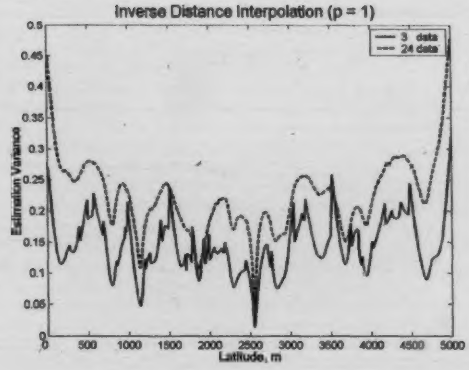
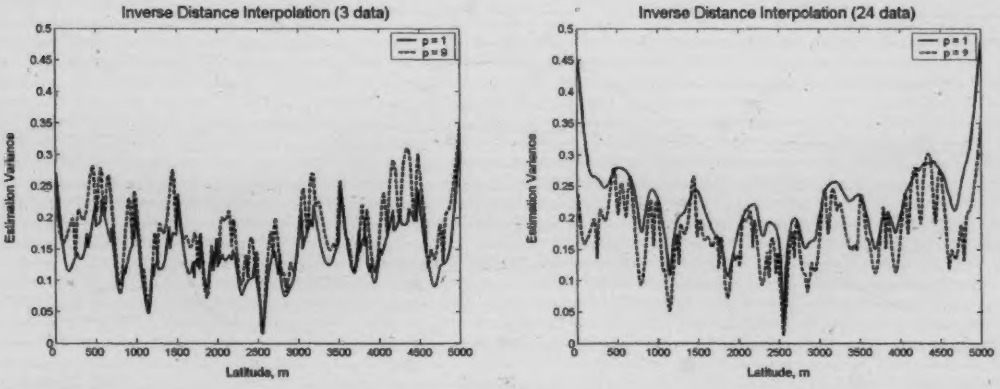


Figure A.5: The estimation variance for the inverse distance interpolator with exponent value of 1 ($p = 1$) as a function of the number of data for the slice at $X = 100$.



a) b)

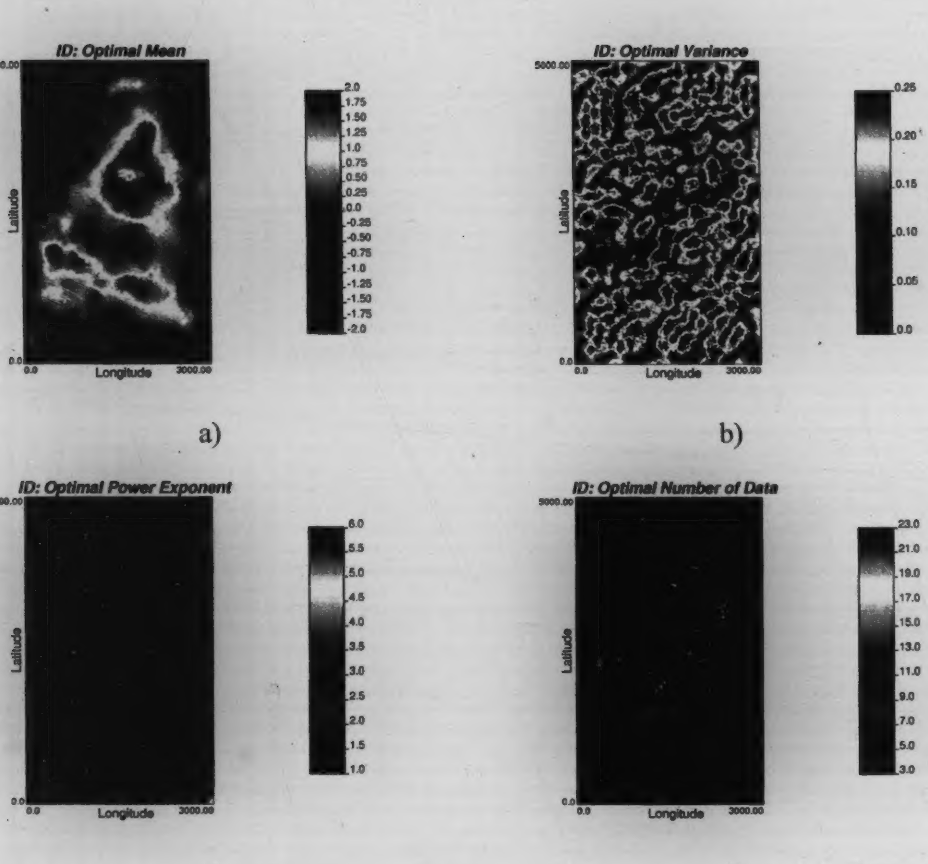
Figure A.6: The estimation variances for the inverse distance interpolator obtained based on 3 data (a) and the estimation variances for the inverse distance interpolator obtained based on 24 data (b) as a function of the power exponent for the slice at $X = 100$.

Figure A.7 shows the result of the optimal local inverse distance interpolation for the mean and variance of the local conditional distributions. The following values for the power exponent were considered: $p = 0.1$ to $p = 12$ with step 0.1; the following values for the number of data were considered: $N = 3$ to $N = 32$ with step 1. Figure A.7 also shows the map of the optimum number of data and exponent power for all estimation locations in the study domain. On average, over the study domain the average number of data used in estimation is 5.7 and an average power exponent is 2.06. However, for some locations the power exponent was as high as 12 while for others as low as 0.1, the same applies to the number of data.

Table A.1 shows results of the cross validation, that is, residual mean square error (MSE) for all 310 data in the study domain obtained based on the inverse distance interpolation with 3, 6, 12 and 24 data and power exponent $p = 1, 2, 3, 4$ and 6. Table A.1 also shows results of the cross validation, that is, MSE obtained based on the local inverse distance interpolation. For comparison it also show residual MSE obtained based on simple kriging and ordinary kriging with 16 data.

Table A.1: Results of the cross validation for all 310 data in the study domain obtained based on the inverse distance interpolation with 3, 6, 12 and 24 data and power exponent $p = 1, 2, 3, 4$ and 6 and based on the local inverse distance interpolation with optimal parameters.

Number of Data	P = 1	P = 2	P = 3	P = 4	P = 6
MSE with 3 data	0.1394	0.1339	0.1326	0.1338	0.1394
MSE with 6 data	0.1676	0.1464	0.1364	0.1330	0.1359
MSE with 12 data	0.1965	0.1618	0.1428	0.1351	0.1358
MSE with 24 data	0.2542	0.1889	0.1529	0.1384	0.1361
MSE with opt. parameters	0.1292				
MSE based on simple kriging with 16 data	0.1143				
MSE based on ordinary kriging with 16 data	0.1129				


ID: Optimal Mean
5000.00
Latitude
0.0
0.0
Longitude
3000.00
2.0
1.75
1.50
1.25
1.0
0.75
0.50
0.25
0.0
-0.25
-0.50
-0.75
-1.0
-1.25
-1.50
-1.75
-2.0
ID: Optimal Variance
5000.00
Latitude
0.0
0.0
Longitude
3000.00
0.25
0.20
0.15
0.10
0.05
0.0
ID: Optimal Power Exponent
5000.00
Latitude
0.0
0.0
Longitude
3000.00
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
ID: Optimal Number of Data
5000.00
Latitude
0.0
0.0
Longitude
3000.00
23.0
21.0
19.0
17.0
15.0
13.0
11.0
9.0
7.0
5.0
3.0

a) b)

c) d)

Figure A.7: Result of the optimal local inverse distance interpolation for the mean (a) and variance (b) of the local conditional distributions; optimal power exponent (c) and optimal number of data (d) for all estimation locations in the study domain.

It is apparent from the considered small example that inverse distance interpolation with locally varying parameters can perform better than inverse distance with constant (global) parameters. Therefore, it would be interesting to analyze the performance of the inverse distance with locally varying parameters in larger number of case studies with different data configurations and different types of spatial variation. Moreover, it would be interesting to compare the performance of the inverse distance with locally varying parameters to simple and ordinary kriging. This analysis will be conducted in the next section.

A.7. Comparison of the Inverse Distance with Locally Varying Parameters with Kriging and Inverse Distance Interpolation

To analyze the performance of inverse distance with locally varying parameters compared to inverse distance interpolation and kriging, the following study was conducted.

Nine exhaustive data sets of 2500 data points each located on 50 by 50 grid with 1m between grid points were selected for analysis. These exhaustive data sets were generated using unconditional LU simulation. Each exhaustive data set exhibits different spatial structure. Spatial structures represent weak, medium and strong correlation between data. Figure A.8 shows the standardized omnidirectional variograms and their theoretical fits for the nine exhaustive data sets. Table A.2 summarizes the variogram models shown in Figure A.8.

Table A.2: Variogram model for the omnidirectional variograms of the nine exhaustive data sets.

Exhaustive Data Set	Variogram Model
Data set 1	$Sph_{a=9.28}(\mathbf{h})$
Data set 2	$Sph_{a=18.5}(\mathbf{h})$
Data set 3	$Sph_{a=39.8}(\mathbf{h})$
Data set 4	$0.29+0.71Sph_{a=9.13}(\mathbf{h})$
Data set 5	$0.33+0.67Sph_{a=21.5}(\mathbf{h})$
Data set 6	$0.27+0.73Sph_{a=40.5}(\mathbf{h})$
Data set 7	$0.55+0.45Sph_{a=10.9}(\mathbf{h})$
Data set 8	$0.62+0.38Sph_{a=20.9}(\mathbf{h})$
Data set 9	$0.72+0.28Sph_{a=37.7}(\mathbf{h})$

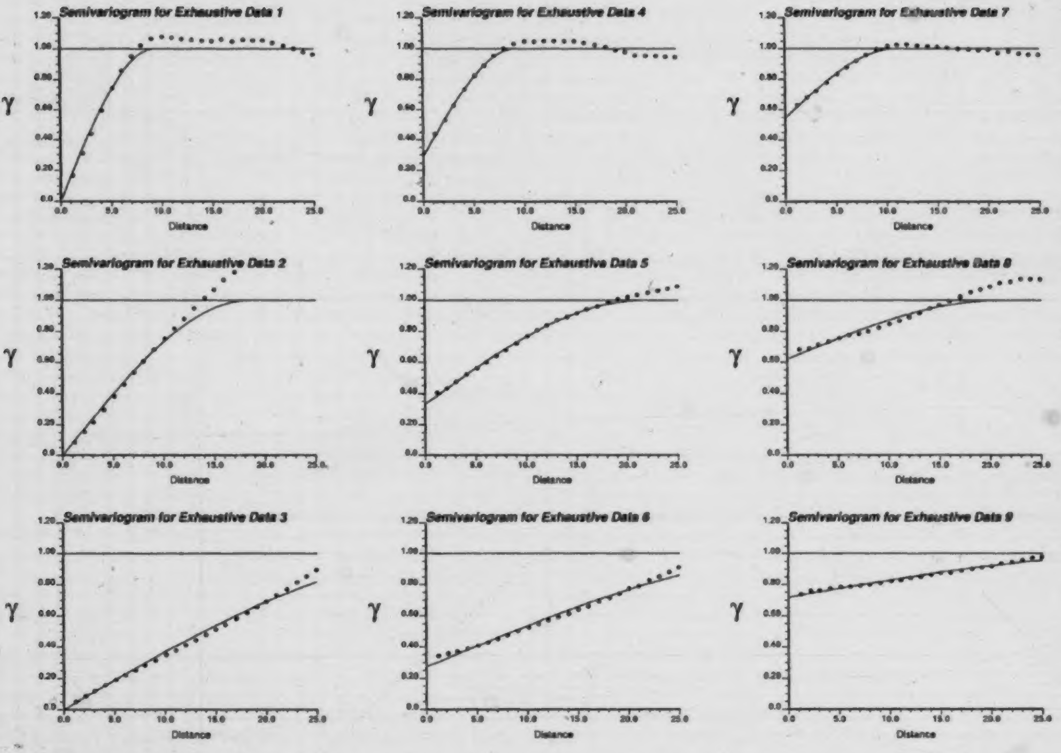
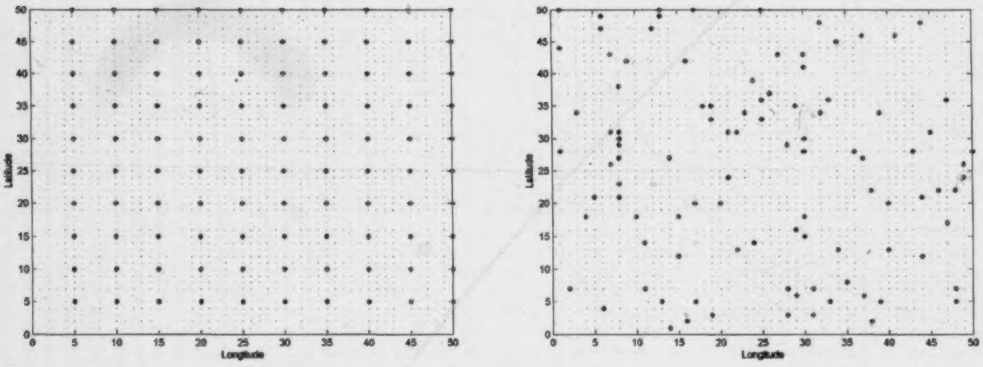


Figure A.8: Standardized omnidirectional variograms and their theoretical fits for the nine exhaustive data sets.

Then from each simulated data set, a data set of size 100 data points on a regular grid (with distance between data of 5m) and a data set of size 100 data points with random clustered pattern were selected (see Figure A.9).

Selected data sets were used for estimation of the study areas of 50 by 50 meters. In estimation four different techniques were used, that is, inverse distance interpolation, ordinary kriging, simple kriging and inverse distance with locally varying parameters. To check the accuracy of the maps produced in estimation all data points except for the estimation data points were used. To evaluate accuracy of different estimation techniques, the mean square error (MSE) criterion was used. In inverse distance interpolation of the study areas of 50 by 50 meters, the following parameters were used: exponent value ranged from 1 to 6, the number of closest samples ranged from 3 to 24.



a) b)

Figure A.9. Locations (asterisk) of 100 points in a) data set selected on a regular grid and in b) data set selected with random clustered pattern. Dots denote other locations with exhaustive data in the study domain of 50 by 50 meters.

In inverse distance interpolation with locally varying parameters, the optimal exponent value was chosen locally from 0.1, 0.2,…,10 and the optimal number of data was also chosen locally from 3, 4,…,24. In kriging of each the test areas, the number of closest samples ranged from 3 to 24. The variogram models for kriging and inverse distance with locally varying parameters were modeled based on selected estimation data points. Example ominidirectional variograms with their fits obtained based on randomly selected 100 data for data set 1 and 100 data on a regular grid for data set 5 and data set 6 are shown in Figure A.10.

Note that the variograms calculated for kriging, in particular, the ones shown in Figure A.10, were fitted using *varfit* program from GSLIB (Deutsch and Journel, 1998). Each fit was obtained based on two spherical variogram structures. The automatic fitting was preferred to manual to avoid any bias from analyzer's side.

Tables A.3 and A.4 show results for the mean square error obtained in estimation of the nine study areas with exhaustive data based on 100 data on a regular grid and 100 randomly selected data, respectively.

Tables A.3: Mean square error obtained in estimation of the nine study areas with exhaustive data based on 100 data on a regular grid.

100 data on regular grid	Best ID	Best ID parameters	Worst ID	SK	OK	LVID	Best Estimator	Worst Estimator
MSE for data 1	0.4195	N=4, p=2	0.7745	0.3907	0.3950	0.4188	SK	ID
MSE for data 2	0.1376	N=4, p=2	0.2722	0.1327	0.1319	0.1353	OK	ID
MSE for data 3	0.0612	N=4, p=2	0.1301	0.0595	0.0586	0.0599	OK	ID
MSE for data 4	0.6650	N=4, p=1	0.8556	0.6413	0.6541	0.6590	SK	ID
MSE for data 5	0.4547	N=5, p=1	0.6195	0.4485	0.4477	0.4476	LVID	ID
MSE for data 6	0.4041	N=16, p=1	0.5721	0.4492	0.4492	0.4506	ID	LVID
MSE for data 7	0.9470	N=9, p=1	1.2548	1.0143	1.0165	1.0099	ID	OK
MSE for data 8	0.7427	N=13, p=1	1.1415	0.7771	0.7669	0.7730	ID	SK
MSE for data 9	0.6995	N=20, p=1	1.3125	0.7347	0.7272	0.7271	ID	SK

Tables A.4: Mean square error obtained in estimation of the nine study areas with exhaustive data based on 100 randomly selected data.

100 randomly selected data	Best ID	Best ID parameters	Worst ID	SK	OK	LVID	Best Estimator	Worst Estimator
MSE for data 1	0.5930	N=24, p=3	0.8107	0.5041	0.5061	0.5675	SK	ID
MSE for data 2	0.1812	N=6, p=2	0.3319	0.1628	0.1627	0.1769	OK	ID
MSE for data 3	0.0765	N=4, p=2	0.1531	0.0638	0.0637	0.0713	OK	ID
MSE for data 4	0.6844	N=21, p=2	0.8847	0.6758	0.6864	0.6689	LVID	OK
MSE for data 5	0.4592	N=8, p=1	0.6510	0.4433	0.4418	0.4549	OK	ID
MSE for data 6	0.4081	N=10, p=1	0.5812	0.3996	0.3964	0.4051	OK	ID
MSE for data 7	0.9290	N=7, p=1	1.2867	0.9826	0.9579	0.9575	ID	SK
MSE for data 8	0.7545	N=15, p=1	1.0685	0.9305	0.7821	0.8067	ID	SK
MSE for data 9	0.7155	N=24, p=1	1.0732	0.8294	0.7314	0.7272	ID	SK

Looking at results of our study, in particular, Tables A.3-A.4, we can conclude the following. When the variogram of the phenomenon under study is closely matched by the experimental variogram calculated based on 100 data points, ordinary and simple kriging outperform both inverse distance interpolation and inverse distance interpolation with locally varying parameters. This is, however, no surprise since kriging provides estimates with global minimum for estimation variance. It is worth noting also that in this case the method proposed in this Appendix A, that is, inverse distance interpolation with locally varying parameters outperforms traditional inverse distance interpolation with any constant (global) values for the number of data and exponent value. The situation, however, becomes more interesting, when the experimental variogram calculated based on sample data differs from the variogram model for the spatial phenomenon under study. Then either kriging techniques or newly proposed approach or inverse distance with correctly chosen global parameters can perform the best depending on the mismatch of the true variogram from the fitted. If the mismatch is significant, inverse distance interpolation with correctly chosen global parameters usually performs best; while kriging, especially simple kriging performs the worst. In this case inverse distance interpolation with locally varying parameters appear to be a very robust technique, that is, it is usually not as strongly affected by the variogram mismatch as simple kriging and even ordinary kriging. Moreover, note that from Tables A.3-A.4 that despite inverse distance interpolation for particularly chosen global parameters in some cases can perform better than inverse distance with locally varying parameters (and kriging), it can be noted that if the parameters for inverse distance interpolation were badly chosen, the results of the inverse distance interpolation with globally chosen parameters could be much worse than respective results of other interpolation techniques. Furthermore note that depending on the data set under study best inverse distance interpolation result is produced by different parameters; no unique value for the number of data and exponent value can produce good results for different data sets in inverse distance interpolation.

A.8. Discussion

There are several reasons why inverse distance interpolation may be preferred over kriging-based techniques. It is simple and applicable to any number of dimensions, it is also robust in estimation, does not suffer from the string effect of kriging (Deutsch, 1993 1994); does not result in negative weights – no screening effect (Deutsch and Journel, 1998); and does not require solving systems of equations for the weights. Moreover, it provides reasonable estimates and is shown in a large number of comparative studies to perform better than kriging-based techniques (Weber and Englund, 1992).

A statistical formalism is proposed for inverse distance estimation. This formalism is based on the assumption of stationarity and a known variogram model. The variance of the inverse distance estimates and the variance of the local conditional distributions are used as measures of accuracy. A general procedure was developed for selecting the optimal number of data and exponent value for the inverse distance estimation of each location separately in the study domain. The developed procedure, referred to as the local inverse distance interpolation was shown to outperform inverse distance interpolation with fixed parameters in the case of closely matched data variogram; and to perform much better than kriging in the case of variogram misspecification.

APPENDIX B

Direct Upscaling of Variograms and Cross Variograms for Scale Consistent Geomodeling

Integration of data from multiple sources and/or multiple scales is a common, yet challenging aspect of geostatistical modeling. Common approaches to data integration are based on a cokriging framework that often assumes the input variogram/covariance models of coregionalization are at a scale consistent with the data and the model grid. The scaling laws for the variogram have often been applied to ensure consistency of the input variogram model; however, these laws are based on a strict assumption of invariance of the variogram shape.

We propose a direct upscaling approach to the variogram that is theoretically derived. The approach (as the scaling laws) is applicable to additive variables. A numerical integration approach to determine the upscaled variogram is presented with an example showing the difference between this theoretical proxy approach and the scaling laws approach. The results show that the shape of the variogram does indeed change with scale. Further, the extension to cross variograms is straightforward and an upscaled consistent linear model of coregionalization is presented.

B.1. Introduction

Data from well cores are among the finest scale of information available for geostatistical inference, and as a result it is often considered to be point scale data with an infinitesimally small volume support. Data from log traces represent an incremental increase in volume support relative to core data; however, these are also considered to be relatively fine scale information. Seismic surveys, if available, generally cover much larger lateral extents with shallow depth. These different data at different support sizes must then be reconciled to some intermediate modeling volume that is determined based on the resolution required and the computational resources available. Integration of these data from various sources and at different volume supports is a longstanding challenge in geostatistical modeling (Kupfersberger et al 1998).

In most cases, the smallest size that the fine scale model can be is limited by computer storage and professional time. Consequently, even the finest scale model is larger than the data support. In the context of estimation, several geostatistical tools exist to facilitate data integration. One set of tools considers that secondary data inform a trend about the primary data; these techniques include external drift and locally varying mean (Marechal 1984; Deutsch and Journel 1998). Another group of methods uses the secondary data as additional conditioning information for estimation of the primary variable; these methods include collocated cokriging (Xu et al 1992; Almeida and Journel 1994), Bayesian updating (Doyen 1996; Ren et al 2007) and block cokriging (Goovaerts 1997). With the exception of block cokriging, most methods assume the primary and all other secondary information are at a consistent scale.

Consistency of scale in the input data and the intended grid can easily be an oversight. In the context of geostatistical simulation, the model is constructed based on a point-support simulation, the input data is considered at a point support and the input variogram corresponds to this same data. This is an acceptably consistent scale model at the point support provided the variable is additive (Journel and Kyriakidis 2004). However, the model is commonly required at an intermediate block scale and high resolution fine scale models are practically infeasible; populating block simulated values is desired.

Direct sequential simulation (DSS) is one possible approach to simulate at a block scale; however, issues with histogram inference and reproduction (Bourgault 1997; Caers 2000; Oz et al 2003) have limited many applications to univariate modeling. Soares (2001) proposed a direct cosimulation approach to handle multivariate problems based on a collocated cokriging approach. If the correlation coefficient between primary and secondary data are properly scaled to account for any scale differentials between the data, then this approach can and will account for different supports.

An alternative to DSS is to consider a 'point' simulation at a block-consistent grid specification, using block averaged data and the block-scale variogram. Once again, the premise for this approach is to ensure consistency in the input data and the required model. Data compositing is a common practice to upscale available hard data. Using this composited data, we can calculate the corresponding composite-support variogram; however, compositing reduces the number of available samples for reliable variogram inference. In such cases, modelers may calculate and fit the variogram using the original support data. This variogram model must then be upscaled for consistency with the model grid and composited data.

Scaling of the variogram model to reflect different volume supports is not new. The scaling laws associated with the variogram are well known (Journel and Huijbregts 1978; Frykman and Deutsch 1999, 2002). They predict how the variogram changes from one volume to another, with specific laws linked to the nugget effect, range and the variance contributions of each constituent structure. A number of simplifying assumptions are made to determine these scaling laws, including volume averaging is performed for non-overlapping volumes, the underlying variable averages linearly, and the shape of the variogram does not change with scale. In particular, the latter assumption of shape invariance of the variogram is a strong and often unrealistic assumption.

This Appendix B proposes to directly upscale the variogram through a numerical integration approach. This numerical approach is used as a proxy to a theoretically developed expression to analytically upscale the variogram for additive varialbes. A review of the variogram scaling laws is provided. The proposed approach to explicitly upscale the variogram model is then described. To facilitate data integration, the extension to cross variograms is straightforward; this is presented in a section that considers the linear model of coregionalization at the required block support. A small example is then used to compare the results of the scaling laws to the direct upscaling approach.

B.2. Regularization

Very seldom, in practice, point data $z(\mathbf{u})$ is available. Most often, the data at hand is defined on a certain support $V = V(\mathbf{u})$ centered on a point $\mathbf{u}$, that is, $z^V(\mathbf{u})$. The value of $z^V(\mathbf{u})$ is the average of the point data $z(\mathbf{u})$ in the volume V, that is,

$$z^V(\mathbf{u}) = \frac{1}{V}\int_V z(\mathbf{w})d\mathbf{w}. \tag{B.1}$$

The value $z^V(\mathbf{u})$ is called the regularization of the point variable $z(\mathbf{w})$ over the volume V (Journel and Huijbregts 1978).

If the point-regionalized variable $z(\mathbf{w})$ is a realization of a second-order stationary random function $Z(\mathbf{w})$, then the regularization of the point random function $Z(\mathbf{w})$ over the volume V is also a second-order stationary random function given by:

$$Z^V(\mathbf{u}) = \frac{1}{V}\int_V Z(\mathbf{w})d\mathbf{w}. \tag{B.2}$$

B.3. Variogram Scaling Laws

The variogram model is linked to the volume support of the data. To represent the volume support that we are interested in, the variogram models need necessarily be scaled. To represent the change in the variogram with the change in the volumetric scale the variogram scaling laws are commonly applied (Frykman and Deutsch 1999, 2002). A short description of these laws follows.

Let us consider a semivariogram model $\gamma_v(\mathbf{h})$ at arbitrary scale v (v usually represents the small core scale):

$$\gamma_v(\mathbf{h}) = C_v^0 + \sum_{i=1}^{k} C_v^i \Gamma_v^i(\mathbf{h}), \tag{B.3}$$

where C_v^0 is the nugget effect, k is the number of nested variogram structures used to fit the experimental variogram of the data, C_v^i, $i = 1,\ldots,k$, is the variance contribution of each nested structure, and $\Gamma_v^i(\mathbf{h})$, $i = 1,\ldots,k$, are individual nested variogram structures with sill of one. Each nested structure is given by analytical function (e.g., spherical, exponential, etc.).

Then a semivariogram model $\gamma_V(\mathbf{h})$ at a larger volume V, where $v \subset V$, is given by

$$\gamma_V(\mathbf{h}) = C_V^0 + \sum_{i=1}^{k} C_V^i \Gamma_V^i(\mathbf{h}), \tag{B.4}$$

where C_V^0 is the nugget effect, C_V^i, $i = 1,\ldots,k$, is the variance contribution of each nested structure, and $\Gamma_V^i(\mathbf{h})$, $i = 1,\ldots,k$, are individual nested variogram structures (all for the scale V). Derivation of the upscaled variogram parameters are then given by (Journel and Huijbregts 1978; Oz and Deutsch 2002):

1. The shape of each individual nested structure (i.e., exponential, spherical) remains invariant when the scale changes. The variogram range of each nested structure increases as the size of the volume V increases. In particular, if a_v^i is the range of $\Gamma_v^i(\mathbf{h})$, $i = 1,\ldots,k$, for the small scale, then the range a_V^i of $\Gamma_V^i(\mathbf{h})$, $i = 1,\ldots,k$, for the large scale is given by

   $$a_V^i = a_v^i + (|V| - |v|), \tag{B.5}$$

   where $|v|$ and $|V|$ relate to the size of the volume in a particular direction.
2. The variance contributions C_v^i, of each nested structure $\Gamma_V^i(\mathbf{h})$, $i = 1,\ldots,k$, change from the small scale to the large scale as follows

   $$C_V^i = C_v^i \frac{1 - \overline{\Gamma}^i(V,V)}{1 - \overline{\Gamma}^i(v,v)}, \tag{B.6}$$

where $\overline{\Gamma}^i(V,V)$ and $\overline{\Gamma}^i(v,v)$ are the average variogram or "gamma-bar" values. In particular, $\overline{\Gamma}^i(V,V)$ represents the mean variogram $\Gamma^i(\mathbf{h})$ when one extremity of the vector **h** describes the domain V and the other extremity of this vector independently describes the same domain. The values of gamma-bar's are usually estimated numerically by volume discretization, that is,

$$\overline{\Gamma}(V,V) = \frac{1}{V}\frac{1}{V}\int_V\int_V \Gamma(\mathbf{y}-\mathbf{y}')d\mathbf{y}d\mathbf{y}' \approx \frac{1}{n}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\Gamma(\mathbf{u}_i-\mathbf{u}_j), \tag{B.7}$$

where n is the number of regular spaced points discretizing the volume V.

3. The variance of the purely random component, called nugget effect, is inversely related to the volume, that is,

$$C_V^0 = C_v^0\frac{|v|}{|V|}. \tag{B.8}$$

It should be noted that as scale increases, the range of correlation increases, and the variogram sill decreases due to high and low values being averaged out. Moreover, it is worth noting also that the scaling laws are established under the following additional assumptions (Journel and Huijbregts 1978):

1. The averaging is performed with non-overlapping volumes.
2. The variables scale in a linear manner.

In general, the assumptions underlying scaling laws appear to be very strong and limiting. In particular, the assumption of no shape change for the variogram nested structures is very unrealistic. It has been observed in many examples that the shape of the experimental variograms at volume V is different from the ones predicted from scaling laws especially for short lag distances (e.g., Frykman and Deutsch 2002). Therefore, a direct approach for exact calculation of the upscaled variograms is of great practical interest.

B.4. Direct Variogram Upscaling

If $Z(\mathbf{w})$ is assumed to be a second-order stationary random function with mean m, covariance $C(\mathbf{h})$ and variogram $2\gamma(\mathbf{h})$. Then $Z^V(\mathbf{u})$ given by

$$Z^V(\mathbf{u}) = \frac{1}{V}\int_V Z(\mathbf{w})d\mathbf{w}$$

is a second-order stationary random function representing the upscaled point random function $Z(\mathbf{w})$ or the regularization of the point random function $Z(\mathbf{w})$ over the volume V centered at location $\mathbf{u}$.

Then, the mean $E(Z^V(\mathbf{u}))$, variance $Var(Z^V(\mathbf{u}))$ semivariogram $\gamma^V(\mathbf{h})$ and covariance $C^V(\mathbf{h})$ of $Z^V(\mathbf{u})$ for the scale V can be calculated based on the mean m, covariance $C(\mathbf{h})$ and semivariogram $\gamma(\mathbf{h})$ of $Z(\mathbf{u})$ for the point scale as follows (Journel and Huijbregts, 1978):

$$\begin{aligned}
&E(Z^V(\mathbf{u})) = m;\\
&Var(Z^V(\mathbf{u})) = C^V(\mathbf{0}) = \overline{C}(V,V);\\
&\gamma^V(\mathbf{h}) = \overline{\gamma}(V,V_{\mathbf{h}}) - \overline{\gamma}(V,V);\\
&C^V(\mathbf{h}) = C^V(\mathbf{0}) - \gamma^V(\mathbf{h}) = \overline{C}(V,V_{\mathbf{h}});
\end{aligned} \tag{B.9}$$

where $V_{\mathbf{h}}$ denote the support V translated from V by the vector $\mathbf{h}$; $\overline{\gamma}(V,V_{\mathbf{h}})$ represents the average of the point semivariogram $\gamma(\mathbf{h})$ when one extremity of the vector $\mathbf{h}$ describes the support V and the other extremity independently describes the translated support $V_{\mathbf{h}}$, that is,

$$\overline{\gamma}(V,V_{\mathbf{h}}) = \frac{1}{V}\frac{1}{V_{\mathbf{h}}}\int_V\int_{V_{\mathbf{h}}}\gamma(\mathbf{w}-\mathbf{x})d\mathbf{w}\,d\mathbf{x} = \frac{1}{V}\frac{1}{V}\int_V\int_V\gamma(\mathbf{w}-(\mathbf{x}+\mathbf{h}))d\mathbf{w}\,d\mathbf{x}; \tag{B.10}$$

$\overline{C}(V,V_{\mathbf{h}})$ represents the average of covariance $C(\mathbf{h})$ and is given by

$$\overline{C}(V,V_{\mathbf{h}}) = \frac{1}{V}\frac{1}{V_{\mathbf{h}}}\int_V\int_{V_{\mathbf{h}}}C(\mathbf{w}-\mathbf{x})d\mathbf{w}\,d\mathbf{x} = \frac{1}{V}\frac{1}{V}\int_V\int_V C(\mathbf{w}-(\mathbf{x}+\mathbf{h}))d\mathbf{w}\,d\mathbf{x}. \tag{B.11}$$

Note that if the semivariogram $\gamma(\mathbf{h})$ of the point-regularized random variable is made up of several nested structures, that is,

$$\gamma(\mathbf{h}) = C^0 + \sum_{i=1}^{k} C^i \gamma^i(\mathbf{h}), \tag{B.12}$$

where C^0 is the nugget effect, k is the number of nested variogram structures, C^i, $i = 1,\ldots,k$, is the variance contribution of each nested structure, and $\gamma^i(\mathbf{h})$, $i = 1,\ldots,k$, are individual nested variogram structures given by analytical function (e.g., spherical, exponential, etc.) with a sill of one. Then, the average of the point semivariogram $\gamma(\mathbf{h})$ can be calculated as

$$\begin{aligned}\bar{\gamma}(V,V_{\mathbf{h}}) &= \frac{1}{V}\frac{1}{V_{\mathbf{h}}}\int_{V}\int_{V_{\mathbf{h}}}\left[C^0 + \sum_{i=1}^{k} C^i \gamma^i((\mathbf{w}-\mathbf{x}))\right] d\mathbf{w}\, d\mathbf{x} \\ &= C^0 + \sum_{i=1}^{k} C^i \left[\frac{1}{V}\frac{1}{V}\int_{V}\int_{V}\gamma(\mathbf{w}-(\mathbf{x}+\mathbf{h}))d\mathbf{w}\, d\mathbf{x}\right] = C^0 + \sum_{i=1}^{k} C^i \bar{\gamma}^i(V,V_{\mathbf{h}}).\end{aligned} \tag{B.13}$$

And, thus, the semivariogram $\gamma^V(\mathbf{h})$ for the scale V is given by

$$\begin{aligned}\gamma^V(\mathbf{h}) = \bar{\gamma}(V,V_{\mathbf{h}}) - \bar{\gamma}(V,V) &= C^0 + \sum_{i=1}^{k} C^i \bar{\gamma}^i(V,V_{\mathbf{h}}) - \left[C^0 + \sum_{i=1}^{k} C^i \bar{\gamma}^i(V,V)\right] \\ &= \sum_{i=1}^{k} C^i \left[\bar{\gamma}^i(V,V_{\mathbf{h}}) - \bar{\gamma}^i(V,V)\right]\end{aligned} \tag{B.14}$$

It is worth noting that semivariogram $\gamma^V(\mathbf{h})$ does not contain any nugget effect. The nugget effect vanishes when upscaling to a larger volume V, it does not matter how large the nugget effect was in the point scale variogram.

Similarly, it can be shown that in this case the covariance for the scale V is given by

$$C^V(\mathbf{h}) = \overline{C}(V,V_{\mathbf{h}}) = C(0) - \left[C^0 + \sum_{i=1}^{k} C^i \bar{\gamma}^i(V,V_{\mathbf{h}})\right]. \tag{B.15}$$

Since the variance $C(0)$ at the point scale for semivariogram $\gamma(\mathbf{h})$ is given by

$$C(0) = C^0 + \sum_{i=1}^{k} C^i, \tag{B.16}$$

then it follows from Equation (B.15) that covariance for the scale V can be calculated as

$$C^V(\mathbf{h}) = \sum_{i=1}^{k} C^i [1 - \bar{\gamma}^i(V,V_{\mathbf{h}})]. \tag{B.17}$$

B.5. Linear Model of Coregionalization (LMC) at a Block Support

In the case when multiple interdependent random variables are available, the spatial relationship between them must be described in a feasible manner. Let us now derive a linear model of coregionalization for N stationary regularizations $\{Z_1^V,\ldots,Z_N^V\}$ of the point random functions $\{Z_1,\ldots,Z_N\}$ over the volume V based on a linear model of coregionalization for $\{Z_1,\ldots,Z_N\}$ at a point support.

Let us consider N stationary point random functions $\{Z_1,\ldots,Z_N\}$. Further let us assume that each point support random function Z_i, $i=1,\ldots,N$, can be expressed as a linear combination of K independent zero mean second-order stationary random functions Y_k, $k=1,\ldots,K$, each with covariance function $C_k(\mathbf{h})$ as follows

$$Z_i(\mathbf{u})=\sum_{k=1}^{K}a_{ik}Y_k(\mathbf{u})+\mu_i. \tag{B.18}$$

The random functions Y_k, $k=1,\ldots,K$, are assumed to be unknown. If we group the random functions Y_k, $k=1,\ldots,K$, according to distinct direct covariances $C_k(\mathbf{h})$, then Equation (B.18) can be rewritten as

$$Z_i(\mathbf{u})=\sum_{l=0}^{L}\sum_{p=1}^{n_l}a_{ip}^{l}Y_p^{l}(\mathbf{u})+\mu_i \tag{B.19}$$

with

$$C(Y_p^l(\mathbf{u}),Y_{p'}^{l'}(\mathbf{u}+\mathbf{h}))=\begin{cases}C^l(\mathbf{h}), & \text{if} \quad p=p' \text{ and } l=l'; \\ 0, & \text{otherwise}; \end{cases} \quad .$$

where $L+1$ is the number of groups with distinct direct covariances; and n_l is the number of random functions with the same covariance $C^l(\mathbf{h})$, $l=0,\ldots,L$.

Then direct and cross covariances between two random variables $Z_i(\mathbf{u})$ and $Z_j(\mathbf{u}+\mathbf{h})$, $i,j=1,\ldots,N$, can be calculated as

$$\begin{aligned} C_{ij}(\mathbf{h}) &= Cov(Z_i(\mathbf{u}), Z_j(\mathbf{u}+\mathbf{h})) = Cov\left(\sum_{l=0}^{L}\sum_{p=1}^{n_l} a_{ip}^{l} Y_p^{l}(\mathbf{u}) + \mu_i, \sum_{l'=0}^{L}\sum_{p'=1}^{n_{l'}} a_{jp'}^{l'} Y_{p'}^{l'}(\mathbf{u}+\mathbf{h}) + \mu_j\right) \\ &= Cov\left(\sum_{l=0}^{L}\sum_{p=1}^{n_l} a_{ip}^{l} Y_p^{l}(\mathbf{u}), \sum_{l'=0}^{L}\sum_{p'=1}^{n_{l'}} a_{jp'}^{l'} Y_{p'}^{l'}(\mathbf{u}+\mathbf{h})\right) = \sum_{l=0}^{L}\sum_{p=1}^{n_l}\sum_{l'=0}^{L}\sum_{p'=1}^{n_{l'}} a_{ip}^{l} a_{jp'}^{l'} Cov(Y_p^{l}(\mathbf{u}), Y_{p'}^{l'}(\mathbf{u}+\mathbf{h})) \\ &= \sum_{l=0}^{L}\sum_{p=1}^{n_l} a_{ip}^{l} a_{jp}^{l} C^{l}(\mathbf{h}). \end{aligned} \qquad (B.20)$$

If we set

$$b_{ij}^{l} = \sum_{p=1}^{n_l} a_{ip}^{l} a_{jp}^{l}, \quad i = 1,\ldots,N, \quad j = 1,\ldots,N, \qquad (B.21)$$

then it follows from Equation (B.20):

$$C_{ij}(\mathbf{h}) = \sum_{l=0}^{L} b_{ij}^{l} C^{l}(\mathbf{h}), \quad i = 1,\ldots,N, \quad j = 1,\ldots,N. \qquad (B.22)$$

Therefore to determine the direct and cross covariances between any two random variables $Z_i(\mathbf{u})$ and $Z_j(\mathbf{u}+\mathbf{h})$, $i, j = 1,\ldots,N$, we need only to determine covariances $C^{l}(\mathbf{h})$, $l = 0,\ldots,L$, and the $(L+1)P^2$ coefficients b_{ij}^{l}. Because a joint matrix of covariance functions $C_{ij}(\mathbf{h})$, $i, j = 1,\ldots,N$, must be positive semi-definite; this requires the covariance models $C^{l}(\mathbf{h})$, $l = 0,\ldots,L$, and L+1 matrices of b_{ij}^{l} coefficients to be positive semi-definite. In practice the covariance models $C^{l}(\mathbf{h})$, $l = 0,\ldots,L$, are chosen to be known positive semi-definite models such as Spherical, Exponential, Gaussian, etc. Model (B.22) is the linear model of coregionalization at a point support.

Moreover, since

$$Z_i^{V}(\mathbf{u}) = \frac{1}{V}\int_{V} Z_i(\mathbf{w})d\mathbf{w}, \qquad (B.23)$$

it follows from Equation (B.19)

$$\begin{aligned} Z_i^{V}(\mathbf{u}) &= \frac{1}{V}\int_{V}\left[\sum_{l=0}^{L}\sum_{p=1}^{n_l} a_{ip}^{l} Y_p^{l}(\mathbf{w}) + \mu_i\right]d\mathbf{w} = \frac{1}{V}\int_{V}\sum_{l=0}^{L}\sum_{p=1}^{n_l} a_{ip}^{l} Y_p^{l}(\mathbf{w})d\mathbf{w} + \frac{1}{V}\int_{V}\mu_i d\mathbf{w} \\ &= \sum_{l=0}^{L}\sum_{p=1}^{n_l} a_{ip}^{l}\left[\frac{1}{V}\int_{V} Y_p^{l}(\mathbf{w})\right]d\mathbf{w} + \mu_i. \end{aligned} \qquad (B.24)$$

Therefore the direct and cross covariances between two regularization random variables $Z_i^{V}(\mathbf{u})$ and $Z_j^{V}(\mathbf{u}+\mathbf{h})$, $i, j = 1,\ldots,N$, are given by

$$
\begin{aligned}
C_{ij}^{V}(\mathbf{h}) &= Cov(Z_i^V(\mathbf{u}), Z_j^V(\mathbf{u}+\mathbf{h}))\\
&= Cov\left(\sum_{l=0}^{L}\sum_{p=1}^{n_l} a_{ip}^{l}\left[\frac{1}{V}\int_V Y_p^l(\mathbf{w})d\mathbf{w}\right]+\mu_i, \sum_{l'=0}^{L}\sum_{p'=1}^{n_{l'}} a_{jp'}^{l}\left[\frac{1}{V}\int_V Y_{p'}^{l'}(\mathbf{y}+\mathbf{h})d\mathbf{y}\right]+\mu_j\right)\\
&= Cov\left(\sum_{l=0}^{L}\sum_{p=1}^{n_l} a_{ip}^{l}\left[\frac{1}{V}\int_V Y_p^l(\mathbf{w})d\mathbf{w}\right], \sum_{l'=0}^{L}\sum_{p'=1}^{n_{l'}} a_{jp'}^{l}\left[\frac{1}{V}\int_V Y_{p'}^{l'}(\mathbf{y}+\mathbf{h})d\mathbf{y}\right]\right)\\
&= \sum_{l=0}^{L}\sum_{p=1}^{n_l}\sum_{l'=0}^{L}\sum_{p'=1}^{n_{l'}} a_{ip}^{l}a_{jp'}^{l'}Cov\left(\frac{1}{V}\int_V Y_p^l(\mathbf{w})d\mathbf{w}, \frac{1}{V}\int_V Y_{p'}^{l'}(\mathbf{y}+\mathbf{h})d\mathbf{y}\right)\\
&= \sum_{l=0}^{L}\sum_{p=1}^{n_l}\sum_{l'=0}^{L}\sum_{p'=1}^{n_{l'}} a_{ip}^{l}a_{jp'}^{l'}\frac{1}{V}\frac{1}{V}\int_V\int_V Cov(Y_p^l(\mathbf{w}), Y_{p'}^{l'}(\mathbf{y}+\mathbf{h}))d\mathbf{w}d\mathbf{y}\\
&= \sum_{l=0}^{L}\sum_{p=1}^{n_l} a_{ip}^{l}a_{jp}^{l}\frac{1}{V}\frac{1}{V}\int_V\int_V C^l(\mathbf{w}-(\mathbf{y}+\mathbf{h}))d\mathbf{w}d\mathbf{y}\\
&= \sum_{l=0}^{L}\sum_{p=1}^{n_l} a_{ip}^{l}a_{jp}^{l}\overline{C}^l(V,V_{\mathbf{h}}).
\end{aligned}
\tag{B.25}
$$

Note that $\overline{C}^l(V,V_{\mathbf{h}})$ represents the average of the point covariance $C^l(\mathbf{h})$ when one extremity of the vector **h** describes the support V and the other extremity independently describes the translated (by vector **h**) support $V_{\mathbf{h}}$.

If we set

$$
b_{ij}^{l} = \sum_{p=1}^{n_l} a_{ip}^{l}a_{jp}^{l}, \quad i=1,\ldots,N, \quad j=1,\ldots,N, \tag{B.26}
$$

then it follows from Equation (B.25):

$$
C_{ij}^{V}(\mathbf{h}) = \sum_{l=0}^{L} b_{ij}^{l}\overline{C}^l(V,V_{\mathbf{h}}),, \quad i=1,\ldots,N, \quad j=1,\ldots,N. \tag{B.27}
$$

Therefore to determine the direct and cross covariances between any two random variables $Z_i^V(\mathbf{u})$ and $Z_j^V(\mathbf{u}+\mathbf{h})$, $i,j=1,\ldots,N$, defined at the block support of size V we need only to determine average covariances $\overline{C}^l(V,V_{\mathbf{h}})$, $l=0,\ldots,L$, and the $(L+1)P^2$ coefficients b_{ij}^l. Because a joint matrix of covariance functions $C_{ij}(\mathbf{h})$, $i,j=1,\ldots,N$, must be positive semi-definite; this requires the covariance models $C^l(\mathbf{h})$, $l=0,\ldots,L$, and L+1 matrices of b_{ij}^l coefficients to be positive semi-definite. Note, however, if the covariance models $C^l(\mathbf{h})$, $l=0,\ldots,L$, are chosen to be known positive semi-definite models such as Spherical, Exponential, Gaussian, etc., then the average covariances calculated based on this models will be also positive-definite (this is a know property of integration). Model (B.27) is the liner model of coregionalization at a block support.

It is interesting to note that if

$$C_{ij}(\mathbf{h}) = \sum_{l=1}^{L} b_{ij}^{l} C^{l}(\mathbf{h}), \quad i = 1,\ldots,N, \quad j = 1,\ldots,N, \tag{B.28}$$

is a feasible linear model of coregionalization at a point support then

$$C_{ij}^{V}(\mathbf{h}) = \sum_{l=1}^{L} b_{ij}^{l} \overline{C}^{l}(V,V_{\mathbf{h}}),, \quad i = 1,\ldots,N, \quad j = 1,\ldots,N. \tag{B.29}$$

is a feasible linear model of coregionalization at a block support and vice versa. This establishes an interesting link between point and block linear model of coregionalization.

B.6. Calculating Average Covariance $\overline{C}(V,V_{\mathbf{h}})$ and Average Variogram $\overline{\gamma}(V,V_{\mathbf{h}})$

Now the only issue remains is to calculate the values of the average covariances $\overline{C}(V,V_{\mathbf{h}})$ and average variogram values $\overline{\gamma}(V,V_{\mathbf{h}})$. Unfortunately, except for simplistic cases (see Journel and Huijbregts 1978), there is no analytical solution for average covariances and semivariograms for any of the commonly used variogram models, that is, spherical, exponential and Gaussian. Interestingly enough, this fact in itself shows that the assumption of invariance of variogram shape to the scale change is unrealistic.

Therefore, to calculate the average covariances $\overline{C}(V,V_{\mathbf{h}})$, giving the covariance $C^{V}(\mathbf{h})$ for the scale V, numerical discretization of the volumes V and $V_{\mathbf{h}}$ is used. Specifically, the average covariances are calculated as

$$\begin{aligned}\overline{C}(V,V_{\mathbf{h}}) = C^{V}(\mathbf{h}) = \frac{1}{V}\frac{1}{V}\int_{V}\int_{V} C(\mathbf{w}-(\mathbf{y}+\mathbf{h}))d\mathbf{w}d\mathbf{y} \approx \\ \frac{1}{n}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} C(\mathbf{u}_{i}-(\mathbf{u}_{j}+\mathbf{h})) = C(0) - \frac{1}{n}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma(\mathbf{u}_{i}-(\mathbf{u}_{j}+\mathbf{h})).\end{aligned} \tag{B.30}$$

The average semivariograms are approximated by

$$\bar{\gamma}(V,V_{\mathbf{h}})=\frac{1}{V}\frac{1}{V}\int_V\int_V\gamma(\mathbf{w}-(\mathbf{x}+\mathbf{h}))d\mathbf{w}\,d\mathbf{x}\approx\frac{1}{n}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma(\mathbf{u}_i-(\mathbf{u}_j+\mathbf{h})), \qquad \text{(B.31)}$$

then the semivariogram $\gamma^V(\mathbf{h})$ for the scale V is calculated as

$$\gamma^V(\mathbf{h})=\bar{\gamma}(V,V_{\mathbf{h}})-\bar{\gamma}(V,V)\approx\frac{1}{n}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma(\mathbf{u}_i-(\mathbf{u}_j+\mathbf{h}))-\frac{1}{n}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma(\mathbf{u}_i-\mathbf{u}_j). \qquad \text{(B.32)}$$

B.7. Example: Scaling Laws vs. Direct Variogram Upscaling

This section is aimed at accentuating the difference between upscaled variograms obtained using scaling laws and the ones obtained from theory via direct variogram upscaling approach.

Let us consider a point second-order stationary random function $Z(\mathbf{w})$ with the following 3D isotropic semivariogram model characterizing its spatial continuity

$$\gamma(\mathbf{h})=0.7Sph_{a=5}(\mathbf{h})+0.3Exp_{a=30}(\mathbf{h});$$

and calculate the semivariogram for the regularization $Z^V(\mathbf{w})$ of the point random function $Z(\mathbf{w})$ over the volume V for different volume sizes V. In particular, we consider three different block sizes for volume V. These are cubes with length 2m, 5m, and 10m.

Figure B.1 shows comparison of the upscaled variograms obtained using scaling laws and direct variogram upscaling approach for the three considered block sizes used for averaging. Figure B.1 also shows point scale variogram.

It can be clearly noted from Figure B.1 that the shape of the variogram changes when changing support of data. With increase in the block size the shape of the block support variograms becomes more pronouncedly Gaussian for small lag distances, see Figure B.2 for a closer examination of the short scale structures observed in Figure B.1.

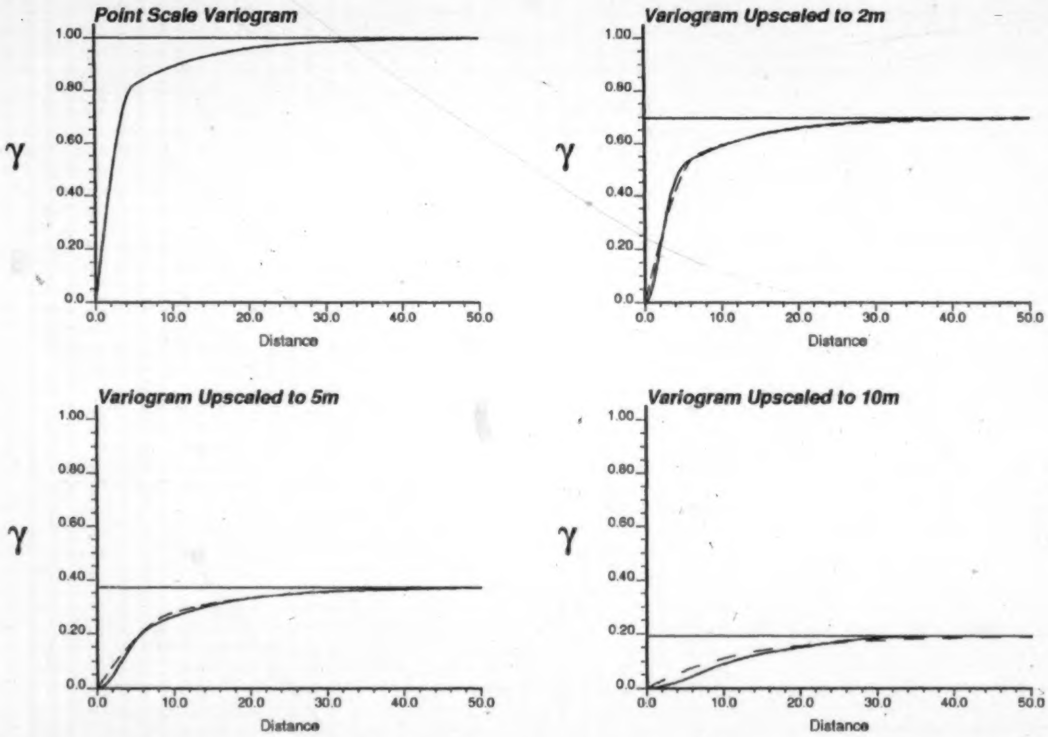


Figure B.1: Comparison of the upscaled variograms obtained using scaling laws (dashed lines) and direct variogram upscaling approach (solid lines) for the block support of 2m, 5m and 10m.

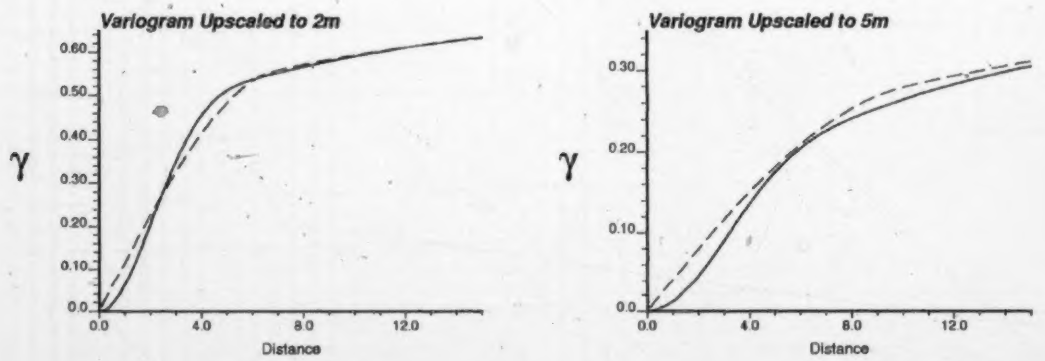


Figure B.2: Comparison of the upscaled variograms obtained using scaling laws (dashed lines) and direct variogram upscaling approach (solid lines) for the block support of 2m and 5m for lag distances up to 10m.

Moreover, the departure between upscaled variogram predicted using scaling laws and theoretical upscaled variograms obtained via direct upscaling approach also increase with increase in the modeling scale (block volume). Therefore, it is apparent that in order to correctly predict the variogram at larger scale using a small scale variogram model a direct upscaling approach should be used; not the scaling laws.

B.8. Discussion

A fundamental assumption underlying the scaling laws for the variogram is that of shape invariance. The theoretically derived approach presented here makes no such assumption. In fact, the examples show that there is a change in the shape of the variogram, specifically a smooth Gaussian structure at short scale can be expected with upscaling to a larger volume. This is consistent with the effects of block averaging.

A practical approach to determine directly upscaled variograms for additive variables is based on a numerical integration that approximates the analytical integral of these variogram models. As with average variogram or average covariance calculations, the approximation is robust given sufficient discretization.

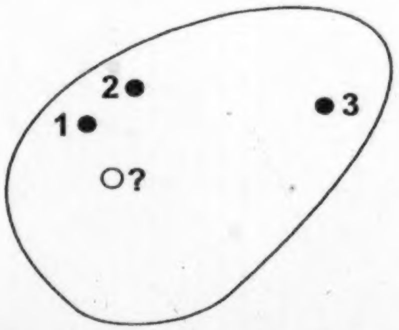
Furthermore, this direct upscaling approach is easily extended to cross-variograms; this facilitates the development of a scale consistent linear model of coregionalization which is required for consistent modeling at the block scale.

This approach presents numerous exciting future research prospects. One area for further development is a method to downscale the block scale variogram, such that fine scale models can be constructed. This is the same objective as the work of Kupfersberger et al (1998), but the goal here will be to avoid use of the scaling laws. Another possible area of research will be to develop a linear model of coregionalization that is consistent at all scales, which could then be used to truly integrate data at different supports without any prior compositing required.

APPENDIX C

Modeling Local Uncertainty accounting for the Uncertainty in Data

Consider the problem of estimation of an unsampled location using surrounding samples, see schematic illustration below.



Standard approach to this problem is kriging. Kriging uses the spatial correlations provided by the variogram to calculate the weights of the sample values surrounding an unsampled location. The weights obtained from the kriging equations minimize the estimation variance and account for the spatial correlation between the surrounding samples and the estimation location (that is, closeness to the estimation location) and between sample themselves (that is, data redundancy). Kriging results in optimal estimation (in the case of a known variogram model) and provides a model for local conditional distributions. In the Gaussian framework, kriging estimate and kriging estimation variance are exactly the mean and variance of the local conditional Gaussian distributions.

Oftentimes, however, the exact sample data are not known due to measurement errors. In this case simple kriging can not be directly applied to infer the local conditional distributions. A theoretical framework for incorporating data uncertainty into calculation of the local uncertainty distributions needs to be developed.

C.1. Simple Kriging

The simple kriging estimator predicts the value of the variable of interest $z(\mathbf{u})$ at the estimation location $\mathbf{u}$ as a linear combination of neighboring observations $z(\mathbf{u}_i), i = 1,\ldots,n(\mathbf{u})$, (Journel and Huijbregts, 1978):

$$z*_{SK}(\mathbf{u}) = \sum_{i=1}^{n(\mathbf{u})} \lambda_i(\mathbf{u}) z(\mathbf{u}_i) + \left[1 - \sum_{i=1}^{n(\mathbf{u})} \lambda_i(\mathbf{u})\right] m, \qquad \text{(C.1)}$$

where m denotes the stationary mean, $\lambda = (\lambda_1(\mathbf{u}),\ldots,\lambda_{n(\mathbf{u})}(\mathbf{u}))^T$ denotes the vector of the simple kriging weights calculated from the normal system of equations

$$\sum_{i=1}^{n(\mathbf{u})} \lambda_i(\mathbf{u}) Cov(z(\mathbf{u}_i), z(\mathbf{u}_j)) = Cov(z(\mathbf{u}), z(\mathbf{u}_j)), \quad j = 1,\cdots,n(\mathbf{u}), \qquad \text{(C.2)}$$

where $Cov(z(\mathbf{u}_i), z(\mathbf{u}_j)), \quad i,j = 1,\cdots,n(\mathbf{u})$, denotes the data-to-data covariance values and $Cov(z(\mathbf{u}), z(\mathbf{u}_j)), \quad j = 1,\cdots,n(\mathbf{u})$, is the data-to-estimation point covariance values. The covariance function is calculated under stationarity through the semivariogram model $\gamma(\mathbf{h})$.

Simple kriging is the best linear unbiased estimator, that is, it provides estimates with minimum error variance $\sigma_{SK}^2(\mathbf{u})$ in the least square sense given by

$$\sigma_{SK}^2(\mathbf{u}) = C(0) - \sum_{i=1}^{n(\mathbf{u})} \lambda_i(\mathbf{u}) Cov(z(\mathbf{u}), z(\mathbf{u}_j)), \qquad \text{(C.3)}$$

where $C(0)$ is the stationary variance.

In the Gaussian framework the local conditional distributions are derived by simple kriging as follows

Uncertainty at the estimation location u is $Z(\mathbf{u}) \sim N(z*_{SK}(\mathbf{u}), \sigma_{SK}^2(\mathbf{u}))$. (C.4)

C.2. Calculating Mean and Variance of the Local Conditional Distributions accounting for the Uncertainty in Data

Let us assume that each of the observations $z(\mathbf{u}_i), i=1,\ldots,n(\mathbf{u})$, available for analysis was measured with some measurement error. Further assume that the measurement errors are distributed according to Gaussian (normal) distribution; thus, uncertainty in each observation (random variable) $Z(\mathbf{u}_i), i=1,\ldots,n(\mathbf{u})$ can be expressed as follows:

$$Z(\mathbf{u}_i) \sim N(\mu_i, \sigma_i^2), \qquad i=1,\ldots,n(\mathbf{u}), \tag{C.5}$$

where μ_i and σ_i^2 denote the mean and variance of the uncertainty distribution in *i*-th data. For now let us assume that the observations $Z(\mathbf{u}_i), i=1,\ldots,n(\mathbf{u})$, represent independent random variables, e.g., each data location was measured using different measurement tool.

When the observations are not longer assumed to be known the mean of the local conditional distributions is a random variable. The variance of the local conditional distributions given in (C.4) is not a random variable, this is because simple kriging variance is homoscedastic, that is, data values independent (see (C.3)). Because the mean of the local conditional distribution is a random variable the uncertainty at the unsampled location **u** is described by the following hierarchical model

$$\begin{gathered} Z(\mathbf{u}) \mid Z^*_{SK}(\mathbf{u}) \sim N(Z^*_{SK}(\mathbf{u}), \sigma^2_{SK}(\mathbf{u})), \\ Z^*_{SK}(\mathbf{u}) \sim N(E[Z^*_{SK}(\mathbf{u})], Var[Z^*_{SK}(\mathbf{u})]), \end{gathered} \tag{C.6}$$

where

$$Z^*_{SK}(\mathbf{u}) = \sum_{i=1}^{n(\mathbf{u})} \lambda_i(\mathbf{u}) Z(\mathbf{u}_i) + \left[1 - \sum_{i=1}^{n(\mathbf{u})} \lambda_i(\mathbf{u})\right] m. \tag{C.7}$$

Note that distribution of $Z^*_{SK}(\mathbf{u})$ is Gaussian because it is a linear combination of Gaussian random variables. Furthermore, due to (C.7), the mean and variance of the distribution for $Z^*_{SK}(\mathbf{u})$ can be calculated as follows:

$$\begin{aligned} E[Z^*_{SK}(\mathbf{u})] &= E\left[\sum_{i=1}^{n(\mathbf{u})}\lambda_i(\mathbf{u})Z(\mathbf{u}_i)+\left[1-\sum_{i=1}^{n(\mathbf{u})}\lambda_i(\mathbf{u})\right]m\right] \\ &= \sum_{i=1}^{n(\mathbf{u})}\lambda_i(\mathbf{u})E[Z(\mathbf{u}_i)]+\left[1-\sum_{i=1}^{n(\mathbf{u})}\lambda_i(\mathbf{u})\right]m = \sum_{i=1}^{n(\mathbf{u})}\lambda_i(\mathbf{u})\mu_i+\left[1-\sum_{i=1}^{n(\mathbf{u})}\lambda_i(\mathbf{u})\right]m = \mu_{Z^*_{SK}(\mathbf{u})}; \end{aligned} \tag{C.8}$$

$$\begin{aligned} Var[Z^*_{SK}(\mathbf{u})] &= Var\left[\sum_{i=1}^{n(\mathbf{u})}\lambda_i(\mathbf{u})Z(\mathbf{u}_i)+\left[1-\sum_{i=1}^{n(\mathbf{u})}\lambda_i(\mathbf{u})\right]m\right] \\ &= Var\left[\sum_{i=1}^{n(\mathbf{u})}\lambda_i(\mathbf{u})Z(\mathbf{u}_i)\right] = \sum_{i=1}^{n(\mathbf{u})}\lambda_i^2(\mathbf{u})Var[Z(\mathbf{u}_i)] = \sum_{i=1}^{n(\mathbf{u})}\lambda_i^2(\mathbf{u})\sigma_i^2 = \sigma^2_{Z^*_{SK}(\mathbf{u})}. \end{aligned} \tag{C.9}$$

Thus, it follows that the local conditional distributions in the case of data uncertainty can be expressed using the following hierarchical model:

$$\begin{aligned} Z(\mathbf{u}) \mid Z^*_{SK}(\mathbf{u}) &\sim N(Z^*_{SK}(\mathbf{u}),\sigma^2_{SK}(\mathbf{u})), \\ Z^*_{SK}(\mathbf{u}) &\sim N(\mu_{Z^*_{SK}(\mathbf{u})},\sigma^2_{Z^*_{SK}(\mathbf{u})}), \end{aligned} \tag{C.10}$$

where $\mu_{Z^*_{SK}(\mathbf{u})}$ and $\sigma^2_{Z^*_{SK}(\mathbf{u})}$) are given in (C.8)-(C.9). Moreover, note that the mean and the variance of local condition distributions is given by

$$E[Z(\mathbf{u})] = E[E[Z(\mathbf{u}) \mid Z^*_{SK}(\mathbf{u})]] = E[Z^*_{SK}(\mathbf{u})] = \mu_{Z^*_{SK}(\mathbf{u})}; \tag{C.11}$$

$$\begin{aligned} Var[Z(\mathbf{u})] &= E[Var[Z(\mathbf{u}) \mid Z^*_{SK}(\mathbf{u})]] + Var[E[Z(\mathbf{u}) \mid Z^*_{SK}(\mathbf{u})]] \\ &= E[\sigma^2_{SK}(\mathbf{u})] + Var[Z^*_{SK}(\mathbf{u})] = \sigma^2_{SK}(\mathbf{u}) + \sigma^2_{Z^*_{SK}(\mathbf{u})}. \end{aligned} \tag{C.12}$$

The shape of the local uncertainty in $Z(\mathbf{u})$ is Gaussian.

It worth noting that when the observations $Z(\mathbf{u}_i), i = 1,\ldots,n(\mathbf{u})$, do not represent independent random variables, but are correlated with a prescribed correlation structure, the mean and variance of the local conditional distributions can be calculated following the same steps as before except variance of $Z^*_{SK}(\mathbf{u})$ needs to be calculated as

$$\begin{aligned} Var[Z^*_{SK}(\mathbf{u})] &= Var\left[\sum_{i=1}^{n(\mathbf{u})}\lambda_i(\mathbf{u})Z(\mathbf{u}_i)+\left[1-\sum_{i=1}^{n(\mathbf{u})}\lambda_i(\mathbf{u})\right]m\right] \\ &= Var\left[\sum_{i=1}^{n(\mathbf{u})}\lambda_i(\mathbf{u})Z(\mathbf{u}_i)\right] = \sum_{i=1}^{n(\mathbf{u})}\sum_{j=1}^{n(\mathbf{u})}\lambda_i(\mathbf{u})\lambda_j(\mathbf{u})Cov[Z(\mathbf{u}_i),Z(\mathbf{u}_j)] = \sigma^2_{Z^*_{SK}(\mathbf{u})}. \end{aligned} \tag{C.13}$$

Moreover note that the above derivations heavily rely on the assumption that the variogram model for the study domain is known; uncertainty in the data does not impact the assumption of the stationary variogram model.

C.3. Small Examples

C.3.1. Example 1

Let us now consider data configuration shown in Figure C.1. In total, there are 4 conditioning data available for inference of the local conditional distribution at the unsampled location. All conditioning data are known subject to measurement errors; the distributions of the conditioning data are Gaussian with different means μ_i and variances σ_i^2, $i = 1,\ldots,4$, see Table C.1 below. Study domain of size 10 by 10 units is assumed to be stationary; stationary mean and variance are 0 and 1, respectively. The variogram of the data is a single structured spherical with nugget effect of 0 and range of correlation of 10 units.

Table C.1: Data locations and values

	Data 1	Data 2	Data 3	Data 4	Unsampled Location
X position	1	5	9	3	5
Y position	3	7	8	2	5
Value	$N(\mu_1,\sigma_1^2)$	$N(\mu_2,\sigma_2^2)$	$N(\mu_3,\sigma_3^2)$	$N(\mu_4,\sigma_4^2)$	?

We will vary the means and variances of the conditional data distributions to assess the impact of data uncertainty on the resulting local uncertainty distribution inferred from simple kriging. First, let us fix μ_i's as follows

$$\mu_1 = 0.8; \quad \mu_2 = 0.2; \quad \mu_3 = -0.4; \quad \mu_4 = -0.1;$$

and examine the effect of σ_i^2's. Table C.2 show results for five different scenarios for σ_i^2's. Note that Table C.2 shows only results for the variance of the local distribution of uncertainty accounting for data uncertainty, that is, $Var[Z(\mathbf{u})]$; this is because the mean of the local conditional distribution is independent of σ_i^2's and equal to 0.0884.

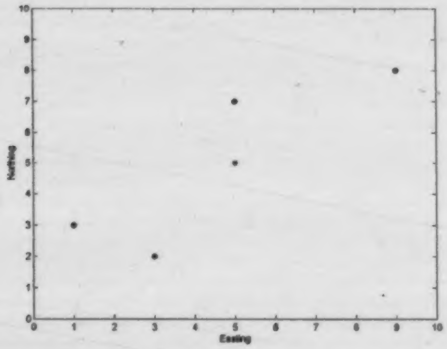


Figure C.1: Data configuration for Example 1.

Table C.2: Effect of σ_i^2's on the local uncertainty distribution.

	Case 1	Case 2	Case 3	Case 4	Case 5
σ_1^2	0	0.1	0.3	0.5	0.8
σ_2^2	0	0.2	0.4	0.6	0.9
σ_3^2	0	0.1	0.2	0.2	0.6
σ_4^2	0	0.3	0.4	0.4	0.7
$Var[Z(\mathbf{u})]$	0.4094 (σ_{SK}^2)	0.5073	0.5871	0.6574	0.7915

It can be clearly noted from Table C.2 that with increase in the data uncertainty (that is, increase in the variance of the conditional data distributions), the variance of the local conditional distribution at the unsampled location increases. Moreover, when there is no uncertainty in the conditional data; the variance of the local conditional distribution at the unsampled location is equal to simple kriging variance.

On the other hand, if we fix σ_i^2's as:

$$\sigma_1^2 = 0.8; \quad \sigma_2^2 = 0.2; \quad \sigma_3^2 = 0.3; \quad \sigma_4^2 = 0.4;$$

we can observe that with increase in the mean of the conditional data distributions, the mean of the local conditional distribution at the unsampled location increases, see Table C.3.

Table C.3: Effect of μ_i's on the local uncertainty distribution.

	Case 1	Case 2	Case 3	Case 4
μ_1	-0.8	-0.2	0.2	1
μ_2	-0.2	0.2	0.2	1
μ_3	-0.4	0.4	0.4	1
μ_4	-0.1	0.1	0.1	1
$E[Z(\mathbf{u})]$	-0.1933	0.1654	0.1765	0.9780

Note that Table C.3 shows only results for the mean of the local distribution of uncertainty accounting for data uncertainty, that is, $E[Z(\mathbf{u})]$; this is because the mean of the local conditional distribution is independent of μ_i's and equal to 0.5176.

It is worth noting that the results shown in Tables C.2-C.3 were theoretically calculated from Equations (C.11)-(C.12) of Section C.2. There is, however, another much more computationally intensive approach based on Monte Carlo simulation to obtain the same result. Specifically, in order to calculate the mean and variance of the local uncertainty distribution accounting for parameter uncertainty via Monte Carlo simulation approach the following steps need to be undertaken:

1. At each of the conditioning data locations draw a value from the conditioning data distribution using Monte Carlo simulation approach;
2. Apply simple kriging to calculate the mean and variance of the local conditional distribution using the conditional data generated in 1;
3. Draw a value from the local conditional distribution obtained in 2. Add to the database;
4. Repeat steps 1-3 many times, say 20000.

To show the equivalence of the theoretically derived local conditional distributions of uncertainty and the ones obtained using Monte Carlo simulation, let us repeat analysis of Table C.2. Results are shown in Tables C.4.

Table C.4: Theoretically-derived approach vs. Monte-Carlo simulation: Variance of the local uncertainty distribution.

$Var[Z(\mathbf{u})]$	Case 1	Case 2	Case 3	Case 4	Case 5
Theory	0.4094	0.5073	0.5871	0.6574	0.7915
Simulation	0.4071	0.5078	0.5884	0.6545	0.7974

The results of theoretically-derived approach vs. Monte-Carlo simulation approach match perfectly; the difference between results of both approaches could have been even smaller if instead of 20000 drawings in Monte-Carlo approach 100000 or more were used.

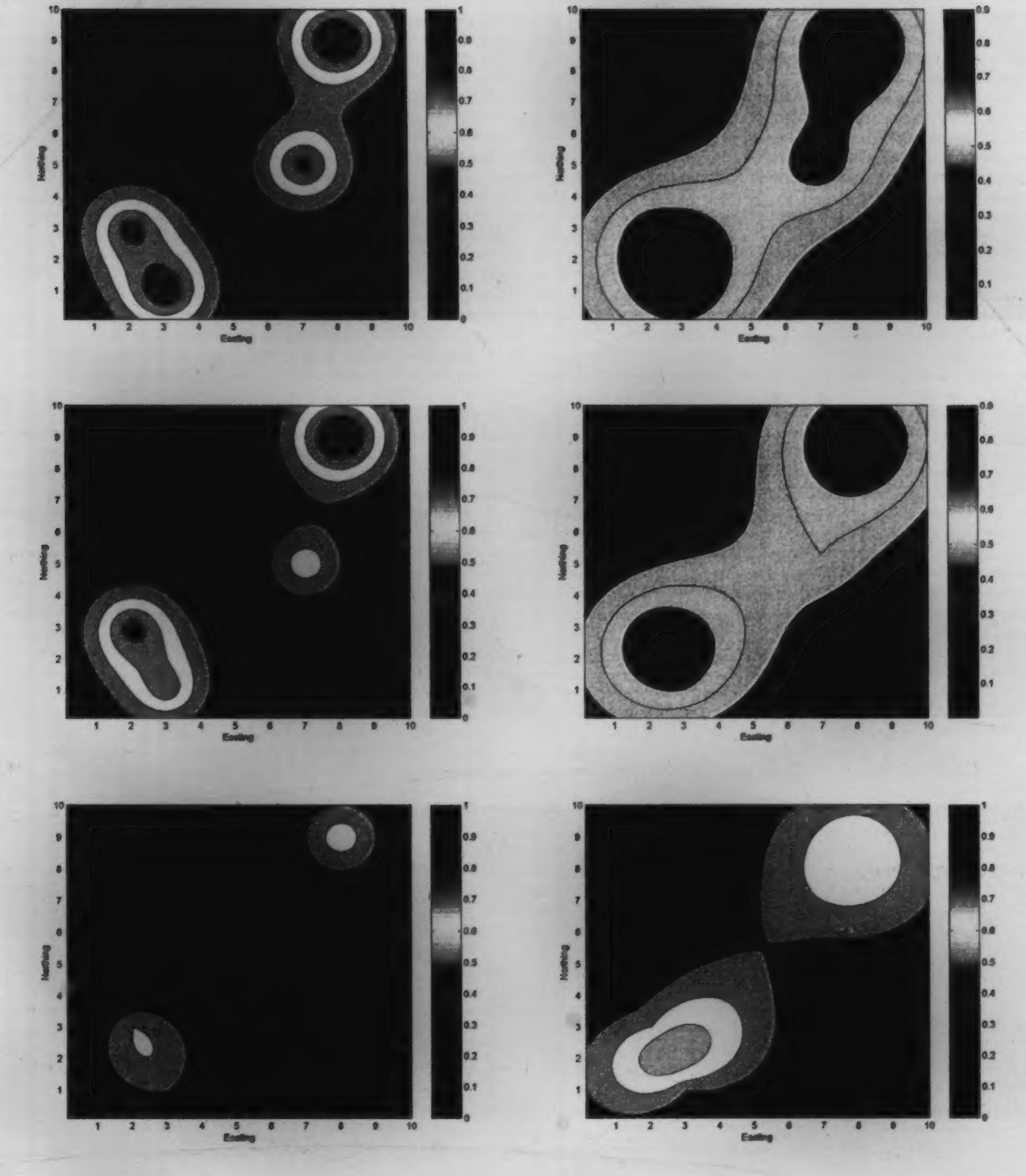
C.3.2. Example 2

To further understand the influence of the data uncertainty on the local conditional distributions at the unsampled locations, let us asses the change in the variance of the local conditional distributions (accounting for data uncertainty) over the study domain. Let us consider the same data configuration as before; set the means of the conditioning data distributions at:

$$\mu_1 = 0.8; \quad \mu_2 = 0.2; \quad \mu_3 = -0.4; \quad \mu_4 = -0.1;$$

and consider three different cases, that is, case 3, case 4 and case 5, for σ_i^2's, see Table C.2. In present study let us also consider two different variogram models, both single structured spherical with nugget effect of 0, but one with range of correlation equal to 10 units and the other one with a range of 5 units and let us compare results.

Figure C.2 shows results obtained in each case. It can be clearly noted from Figure C.2 that with increase in the range of continuity, the variance of the local conditional distributions decreases. The variance of the local conditional distributions usually lies in the interval from 0 to 1. However, it can be also higher than 1, see Table C.5.



a) b)

Figure C.2: Variance of the local conditional distributions accounting for the uncertainty in the data obtained based on a single structured spherical variogram with nugget effect of 0 and range of continuity 5 (a) and 10 (b) : case 3 (top), case 4 (middle) and case 5 (bottom).

Table C.5: Maximum variance f the local conditional distributions over the study domain.

Maximum $Var[Z(\mathbf{u})]$	Case 3	Case 4	Case 5
Range 5	1	1	1.0021
Range 10	0.9967	0.9993	1.0446

C.4. Discussion

In this Appendix C a new interesting framework for incorporation of the data uncertainty into geostatistical estimation was approach. The theory behind the methodology was developed in detail; theoretical results were compared with practical results obtained via direct Monte Carlo simulation. Two small examples illustrating the change in the local uncertainty when incorporating data uncertainty were presented.